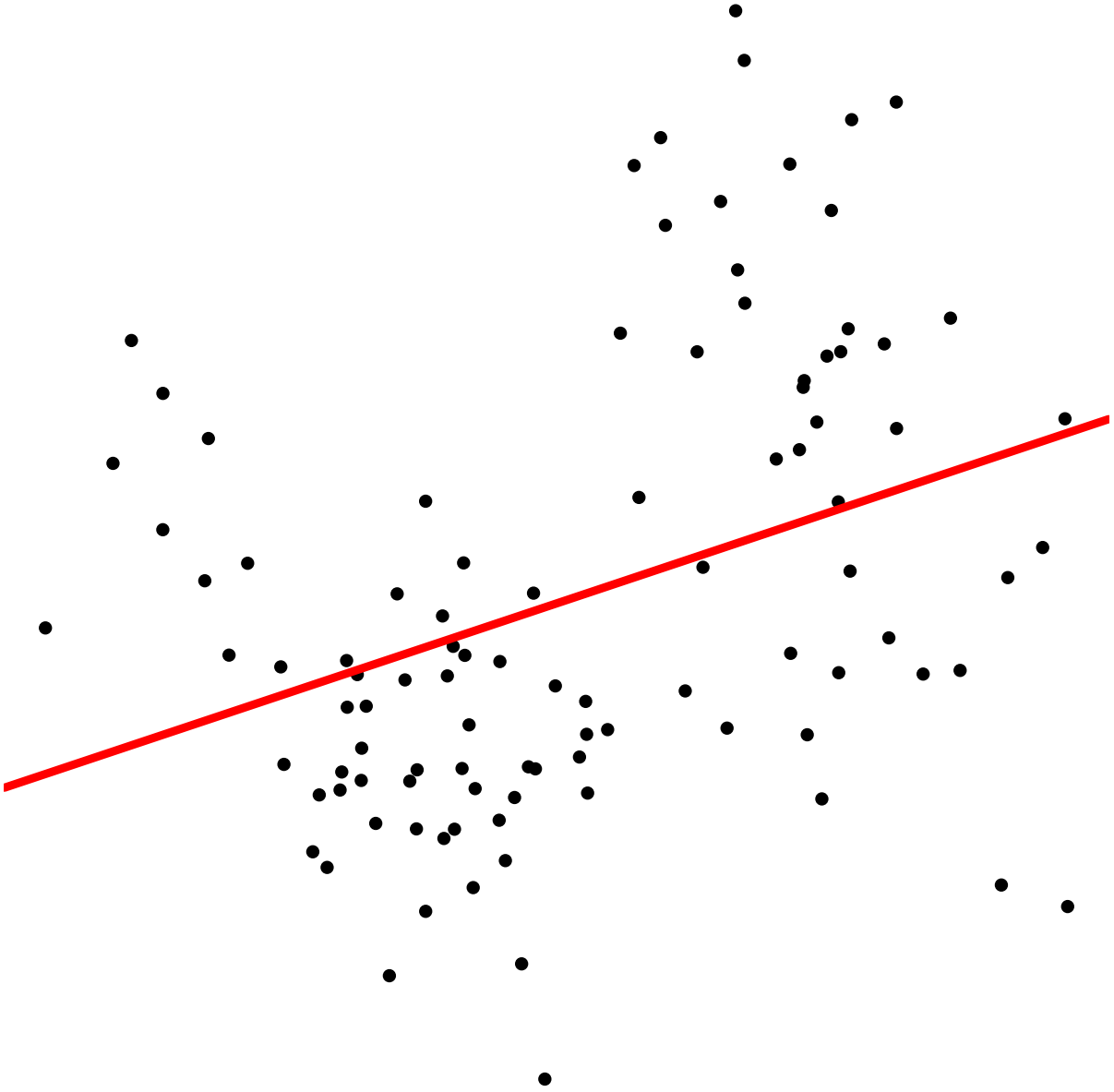
Theory of Agnostic Statistics

Peter M. Aronow and Benjamin T. Miller

September 6, 2015



For Our Parents

Contents

I do not pretend to know what I do not know.

— SOCRATES

0 Introduction

Humans are allergic to change. They love to say, "We've always done it this way." I try to fight that. That's why I have a clock on my wall that runs counter-clockwise.

— GRACE HOPPER

It is possible to make statistical claims under assumptions that researchers in the social and health sciences would find credible. In fact, it is possible to do so using simple and common statistical methods.

Yet the theory of statistics is often made opaque by an unnecessary focus on the construction of abstract and simplistic models of the process that gives rise to the researcher's data. The operating characteristics of statistical methods are then typically assessed under the assumption that the motivating model holds. It is no surprise, then, that in recent years we have seen a proliferation of new methods, usually purporting to solve some problem by invoking alternative modeling assumptions. These motivating assumptions are usually far stricter than we are willing to tolerate on substantive grounds. The result amounts to "cargo cult" statistical inference: we may use the methods that statisticians use, but we do not fully appreciate that the implied models do not approximate the processes under study.

In this book, we take a different approach: we consider what can be learned *agnostically*—that is, without imposing a restrictive model of the data-generating process. Assuming that the data the researcher collects are produced by repeated draws from some random generative process, we can learn about some of the characteristics of the process that generated them without any further assumptions. We can estimate a feature of this random generative process (e.g., "Our guess about the average height in the population from which we are drawing random samples is 5.6 feet."), and we can make probabilistic statements describing the uncertainty of our estimates (e.g., "We can state with 95% confidence that the average height in this population lies between 5.3 feet and 5.9 feet."). Simple statistical methods for estimation and inference, including standard tools such as linear regression, allow us to approximate these features without making any stronger assumptions.

We then ask: under what circumstances do these features have a substantive, or indeed causal, interpretation? This, of course, necessitates detailed knowledge of the process at hand. We will discuss assumptions with clear substantive interpretations that allow us to generalize from the statistical model to broader phenomena, including missing data and causal inference. These assumptions are strong, but they can be justified with a convincing research design.

We will not eschew models entirely in this book—indeed, the construct of probability is itself a model. But, insofar as we use models, we will use them to approximate a more complex inferential target. Restrictive models may help to give us a better estimate in any given sample, but we must be careful to interpret the results using the assumptions that we believe. This is the essence of agnostic statistics: we understand our statistical methods not in terms of a particular set of assumptions that motivate them, but rather in terms of their operating characteristics even when these assumptions fail.

Our goal in this book is to allow readers not just to be competent practitioners, but also to help readers develop a deep understanding of how statistical methods really work. Once these ideas are internalized, readers will be able to critically evaluate the value and credibility of both applied work and statistical

methods. There is no magic in statistics, and it is certainly possible for the motivated researcher to conduct cutting-edge research by working from first principles. No amount of fancy new methods or statistical razzle-dazzle can substitute for a solid research design and a sound understanding of the problem at hand.

0.1 Expectations

This book is designed to be used in conjunction with a one-year sequence in statistics for graduate students or advanced undergraduates in the social or health sciences, but we recommend it for all applied researchers who seek to internalize an agnostic interpretation of statistics.

While this book is conceptually complex, the mathematics are relatively simple. We expect that the readers of this book will have had some exposure to the ideas of probability and statistics at the undergraduate level; while not required, it will significantly ease the readers' experience with the book. Some mild calculus will be used: nothing much beyond partial derivatives and integrals, and even then numerical methods will typically suffice for anything more complicated than a polynomial. Some basic set theory will also be required for our exposition of probability theory. Notably, we avoid the gratuitous use of linear algebra. Concepts from more advanced areas of mathematics (e.g., measure theory) appear in some technical footnotes, but these can be safely ignored by readers not familiar with these subjects. We try to include the proofs of as many of the theorems and principles in this book as possible, though we shall find it necessary to omit those that require advanced mathematics not covered in this book.

0.2 What We Will Not Cover

Although we will cover many of the fundamental topics in statistical inference, there is much that will not be covered in this book. We discuss here some notable exclusions. We will only operate under the frequentist paradigm, omitting the Bayesian mode of inference. We will not detail design-based inference, i.e., inference explicitly predicated on the act of randomization. We will not cover statistical modeling in the traditional sense, including model testing, model diagnostics, and graph-based models. Our discussion of causal inference will largely be limited to a core, fundamental case: cross-sectional data with a single treatment. Many recent and popular strategies in causal inference (e.g., instrumental variables, regression discontinuity designs, inference from longitudinal data) will not be detailed here. Our treatment of penalized regression techniques will be very brief and cursory. Although we view these topics as important, they are ultimately beyond the scope of our book.

0.3 Acknowledgements

We thank the following for valuable conversations and feedback on this book: Ellen Alpert, Jonathon Baron, Tommaso Bardelli, Kassandra Birchler, Xiaoxuan Cai, Forrest Crawford, Germán Feierherd, Don Green, Anand Gupta, Josh Kalla, Mary McGrath, Betsy Levy Paluck, Joel Middleton, Molly Offer-Westort, Lilla Orr, Kyle Peyton, Thomas Richardson, Cyrus Samii, and Pavita Singh. We owe a special

debt of gratitude to Winston Lin, whose comments on early drafts of this book were extraordinarily helpful. We also thank our classes, the Yale students in PLSC 500 and PLSC 503, who suffered through early iterations of the manuscript.

As usual, we are standing on the shoulders of giants. Our book owes enormously to prior work, most notably Goldberger's *A Course in Econometrics*, particularly in our treatment of linear regression. Other notable influences include Wasserman's *All of Statistics* and *All of Nonparametric Statistics*, Angrist and Pischke's *Mostly Harmless Econometrics*, and Manski's *Partial Identification of Probability Distributions*. Our text should not be considered a substitute for these important works, but rather a complement.

1 Probability Theory

Though there be no such thing as chance in the world, our ignorance of the real cause of any event has the same influence on the understanding.

— DAVID HUME

In this chapter, we provide a formalization of probability theory, and establish some basic theorems that will allow us to formally describe random generative processes and quantify relevant features of these processes. We believe that it is important for researchers to understand the assumptions embedded in mathematical probability theory before attempting to make statistical claims. Our approach is thus somewhat unconventional, in that we focus on describing random variables *before* we consider data, so as to have well-defined inferential targets.[1]

Many of the ideas in this chapter are essential for understanding the key ideas in this book, and our treatment of them may be somewhat unfamiliar for readers whose prior training is in applied econometrics or data analysis. We therefore recommend that even relatively statistically sophisticated readers (and also readers otherwise uninterested in probability theory) read the content of this chapter, as its presentation will inform our discussion of more advanced and applied topics in subsequent chapters.

As this is not primarily a book on probability theory, our treatment of this rich area of mathematics is necessarily abbreviated. Mathematically sophisticated readers (i.e., those comfortable with the concepts and notation of calculus, basic set theory, and proofs) should have little difficulty learning the essentials of probability theory from this chapter. For readers who have already been exposed to mathematical probability theory, this chapter should serve as a review of the concepts that will be important for the rest of this book. For readers with neither previous exposure to probability theory nor fluency in college-level mathematics, we strongly recommend consulting a textbook on this subject.[2] As we proceed, some mathematical notation will not be defined in the main text; however, a mathematical glossary is included in Appendix A.

1.1 Random Events

Probability theory is a *mathematical construct* used to represent processes involving randomness, unpredictability, or intrinsic uncertainty. We shall refer to such phenomena as *random generative processes*. In this section, we present the basic principles of probability theory used to describe random generative processes.

[1] This is sometimes referred to as the "population first" approach (Angrist & Pischke 2009), for reasons that will become clear in Chapter 2.

[2] We recommend Chapter 1 of Wasserman (2004) for a concise treatment, though we are fond of Part IV of Freedman, Pisani, and Purves (1998) as a very friendly introduction to the basics of probability theory. For a more thorough treatment of mathematical probability theory, we recommend Chapters 2-5 of Wackerly, Mendenhall, and Scheaffer (2002).

1.1.1 Fundamentals of Probability Theory

We begin by introducing the concepts and notation used to represent the basic elements of random generative processes. We can think of a random generative process as a mechanism that selects an *outcome* from among multiple possible outcomes. This mechanism could be flipping a coin or rolling a die, drawing a ball from an urn, selecting a person at random from a group of people, or any other process in which the outcome is in some sense uncertain. A single instance of selecting an outcome is known as a *draw* from or *realization* of the generative process.[3]

For a given generative process, we let Ω be the set of all possible outcomes, called the *sample space*. Individual outcomes (sometimes known as *sample points*) are denoted by $\omega \in \Omega$. Outcomes can be represented by numbers, letters, words, or other symbols—whatever is most convenient for describing every distinct possible outcome of the random generative process. For example, if we wanted to describe a single roll of a six-sided die, we could let $\Omega = \{1, 2, 3, 4, 5, 6\}$. To describe a roll of two six-sided dice, we could let Ω be the set of all ordered pairs of integers between 1 and 6, i.e., $\Omega = \{(x, y) \in \mathbb{Z}^2 : 1 \leq x \leq 6, 1 \leq y \leq 6\}$. To describe a fair coin flip, we could let $\Omega = \{Heads, Tails\}$ or $\Omega = \{H, T\}$. To describe choosing a random person in the United States and measuring their height in inches, we could let Ω be the set of all positive real numbers, $\Omega = \mathbb{R}^+$. And so on.

An essential concept in probability theory is the idea of a *random event*. Events are subsets of Ω, which are denoted by capital Roman letters, e.g., $A \subseteq \Omega$. Whereas Ω describes all distinguishable states of the world that could result from the generative process, an event may occur in multiple states of the world, so we represent it as a set containing all states of the world in which it occurs. For example, in the case of rolling a single six-sided die, we could represent the event of rolling an even number by the set $A = \{\omega \in \Omega : \omega \text{ is even}\} = \{2, 4, 6\}$. Of course, an event can also correspond to a single state of the world, e.g., the event of rolling a 3, which we might represent by the set $B = \{3\}$. Such events are variously known as *atomic events*, *elementary events*, or *simple events*. A set of events for a random generative process is called an *event space* if it satisfies certain properties.

Definition 1.1.1. *Event Space*

A set S of subsets of Ω is an event space if it satisfies the following:

- *Nonempty: $S \neq \emptyset$.*
- *Closed under complements: if $A \in S$, then $A^C \in S$.*
- *Closed under countable unions: if $A_1, A_2, A_3, ... \in S$, then $A_1 \cup A_2 \cup A_3 \cup ... \in S$.*[4]

[3] The term *experiment* is also commonly used, but we shall refrain from this usage to avoid confusion with experiments in the ordinary sense of the term.

[4] A set S of subsets of another set Ω that satisfies these properties is formally known as a *σ-algebra* or *σ-field* on Ω. Some readers well-versed in set theory may wonder why we do not simply let $S = \mathcal{P}(\Omega)$, the power set (i.e., set of all subsets) of Ω. For reasons that we will not discuss in this book, this does not always work; for some sample spaces Ω, it is impossible to define the probability of every subset of Ω in a manner consistent with the axioms of probability (see Definition 1.1.2). We need not worry too much about this point, though; in practice we will be able to define the probability of any event of interest without much difficulty.

The final component needed to mathematically describe a random generative process is a *probability measure*. A probability measure is a function $P : S \to \mathbb{R}$ that assigns a probability to every event in the event space. To ensure that P assigns probabilities to events in a manner that is coherent and in accord with basic intuitions about probabilities, we must place some conditions on P. Such conditions are provided by the *Kolmogorov probability axioms*, which serve as the foundation of probability theory.[5] These axioms define a *probability space*, a construct that both accords with basic intuitions about probabilities and lends itself to rigorous and useful math.

Definition 1.1.2. *Kolmogorov Axioms*

Let Ω be a sample space, S be an event space, and P be a probability measure. Then (Ω, S, P) is a probability space if it satisfies the following:

- *Non-negativity: $\forall A \in S$, $P(A) \geq 0$, where $P(A)$ is finite and real.*
- *Unitarity: $P(\Omega) = 1$.*
- *Countable additivity: if $A_1, A_2, A_3, ... \in S$ are pairwise disjoint,[6] then*

$$P(A_1 \cup A_2 \cup A_3 \cup ...) = P(A_1) + P(A_2) + P(A_3) + ... = \sum_i P(A_i).$$

We can represent any random generative process by a probability space (Ω, S, P) in a fairly straightforward manner, as illustrated by the following simple example.

Example 1.1.1. *A Fair Coin Flip*

Consider a fair coin flip. Let H represent the outcome "heads" and T represent the outcome "tails." Let $\Omega = \{H, T\}$ and $S = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$. Then

$$P(A) = \begin{cases} 0 & : A = \emptyset \\ \frac{1}{2} & : A = \{H\} \text{ or } A = \{T\} \\ 1 & : A = \{H, T\} \end{cases}$$

The reader can verify that S is a proper event space (i.e., is nonempty and closed under complements and countable unions) and that P satisfies the Kolmogorov axioms, so (Ω, S, P) is a probability space.

Several useful properties of probability follow directly from the Kolmogorov axioms.

[5] This is not the only way to formalize probability. Cox's Theorem, favored by some Bayesians, yields the same results. See, e.g., Jaynes (2003). However, the Kolmogorov axioms are the most common formalization, and the one we will be working with in this book.

[6] Recall that sets A and B are disjoint if $A \cap B = \emptyset$. We say that $A_1, A_2, A_3, ...$ are pairwise disjoint if each of them is disjoint from every other, i.e., $\forall i \neq j$, $A_i \cap A_j = \emptyset$.

Theorem 1.1.1. *Basic Properties of Probability*

Let (Ω, S, P) *be a probability space.*[7] *Then:*

- *Monotonicity:* $\forall A, B \in S$*, if* $A \subseteq B$*, then* $P(A) \leq P(B)$.
- *Subtraction rule:* $\forall A, B \in S$*, if* $A \subseteq B$*, then* $P(B \backslash A) = P(B) - P(A)$.
- *Zero probability of the empty set:* $P(\emptyset) = 0$.
- *Probability bound:* $\forall A \in S$, $0 \leq P(A) \leq 1$.
- *Complement rule:* $\forall A \in S$, $P(A^C) = 1 - P(A)$.

Proof: Let $A, B \in S$ with $A \subseteq B$. Since $B = A \cup (B \backslash A)$, and A and $(B \backslash A)$ are disjoint, countable additivity implies

$$P(B) = P(A) + P(B \backslash A).$$

Non-negative probabilities then imply monotonicity:

$$P(A) = P(B) - P(B \backslash A) \leq P(B).$$

Rearranging this equation yields the subtraction rule:

$$P(B \backslash A) = P(B) - P(A).$$

The subtraction rule, in turn, implies zero probability of the empty set: $A \subseteq A$, so

$$P(\emptyset) = P(A \backslash A) = P(A) - P(A) = 0.$$

Monotonicity and unitarity (and non-negativity) imply the probability bound: since $A \subseteq \Omega$,

$$0 \leq P(A) \leq P(\Omega) = 1.$$

Finally, the subtraction rule and unitarity imply the complement rule:

$$P(A^C) = P(\Omega \backslash A) = P(\Omega) - P(A) = 1 - P(A). \ \square$$

1.1.2 Frequentist Interpretation of Probability

The probability measure of an event describes the proportion of times that event can be expected to occur among many realizations of a random generative process. This interpretation of probability is known as *frequentist probability* or *frequentism*: the probability of an event A is interpreted as representing how

[7] This assumption shall henceforth be implicit in all definitions and theorems referring to Ω, S, and/or P.

frequently A would occur among many, many draws from a random generative process. It is the long-run average or limiting value of the frequency of observing event A among repeated draws from (Ω, S, P).[8]

Probability theory is a *model*, an approximation of reality. Everyday macrophysical processes are not actually characterized by fundamental randomness. Consider, again, the example of a coin flip. In principle, if we could know the exact mass, shape, and position of the coin at the moment it was flipped and the exact magnitude and direction of the force imparted to it by the flipper and of all other forces acting on it, we could predict *with certainty* whether it would land on heads or tails.[9]

The mathematical construct of randomness is a *modeling assumption*, not necessarily a fundamental feature of reality.[10] It allows us to model the outcomes of the coin flip given our uncertainty about the exact nature of the forces that will act on the coin in any particular instance. Similarly, in the social and health sciences, the assumption of randomness allows us to model various outcomes that we might care about, given our uncertainty about the precise features of complex social or biological processes.

1.1.3 Joint and Conditional Probabilities

We often want to describe how the probability of one event relates to the probability of another. The *joint probability* of events A and B is the probability that events A and B will both occur in a single draw from (Ω, S, P).

Definition 1.1.3. *Joint Probability*

For $A, B \in S$, the joint probability of A and B is $P(A \cap B)$.

In words, the joint probability of two events A and B is the probability of the intersection of A and B (which is itself an event in S), i.e., the set of all states of the world in which both A and B occur. We illustrate this point with the following example.

Example 1.1.2. *A Fair Die Roll*

Consider a roll of one fair (six-sided) die. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, $S = \mathcal{P}(\Omega)$ (the power set—i.e., set of all subsets—of Ω), and $P(A) = \frac{1}{6}|A|$, $\forall A \in S$. Let $A = \{\omega \in \Omega : \omega \geq 4\} = \{4, 5, 6\}$ and $B = \{\omega \in \Omega : \omega \text{ is even}\} = \{2, 4, 6\}$. Then

$$P(A \cap B) = P(\{4, 5, 6\} \cap \{2, 4, 6\}) = P(\{4, 6\}) = \frac{1}{6}|\{4, 6\}| = \frac{2}{6} = \frac{1}{3}.$$

Just as $P(A \cap B)$ is the probability that both A *and* B will occur in a single draw from (Ω, S, P), $P(A \cup B)$ is the probability that A *or* B (or both) will occur in a single draw from (Ω, S, P). The following theorem relates these two probabilities.

[8] There are other interpretations of probability, most notably the *Bayesian* interpretation, which treats probability as representing a degree of belief or confidence in a proposition. We do not discuss these alternative interpretations and will be operating under the frequentist paradigm throughout this book.

[9] See Diaconis, Holmes, & Montgomery (2007).

[10] Current thinking in physics suggests that randomness *is* a fundamental feature of quantum-mechanical processes, rather than merely a representation of unknown underlying factors determining individual outcomes. We are not considering quantum mechanics in this book. Suffice it to say that quantum randomness is probably not relevant to the social or health sciences.

Theorem 1.1.2. *Addition Rule*

For $A, B \in S$,

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

Proof: Note that $(A \backslash B)$, $(B \backslash A)$, and $(A \cap B)$ are pairwise disjoint and $(A \cup B) = (A \backslash B) \cup (B \backslash A) \cup (A \cap B)$, so by countable additivity,

$$\begin{aligned} P(A \cup B) &= P(A \backslash B) + P(B \backslash A) + P(A \cap B) \\ &= P\left(A \backslash (A \cap B)\right) + P\left(B \backslash (A \cap B)\right) + P(A \cap B) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B), \end{aligned}$$

where the second step holds because $A \backslash (A \cap B) = A \cap (A \cap B)^C = A \cap (A^C \cup B^C) = (A \cap A^C) \cup (A \cap B^C) = \emptyset \cup (A \backslash B) = A \backslash B$, and likewise $B \backslash (A \cap B) = B \backslash A$, and the third step follows from the subtraction rule, since $A \cap B \subseteq A$ and $A \cap B \subseteq B$. $\square$

In other words, the probability of *either* of two events occurring is equal to the sum of the probabilities of *each* occurring minus the probability of *both* occurring. This naturally yields the following corollary.

Corollary: *If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$.*

We also frequently want to describe the probability of observing event A *given* that we observe event B. This is known as *conditional probability*.

Definition 1.1.4. *Conditional Probability*

For $A, B \in S$ with $P(B) > 0$, the conditional probability of A given B is

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

We can rearrange this definition to obtain another useful formula, the *Multiplicative Law of Probability*.

Theorem 1.1.3. *Multiplicative Law of Probability*

For $A, B \in S$ with $P(B) > 0$,

$$P(A|B)P(B) = P(A \cap B).$$

Proof: Simply rearrange Definition 1.1.4. $\square$

One of the most important theorems regarding conditional probability is *Bayes' Rule* (also known as *Bayes' Theorem* or *Bayes' Law*).

Theorem 1.1.4. *Bayes' Rule*

For $A, B \in S$ with $P(A) > 0$ and $P(B) > 0$,

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

Proof: By the Multiplicative Law of Probability, $P(A \cap B) = P(B|A)P(A)$. So by the definition of conditional probability,

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

The following example illustrates the above concepts.

Example 1.1.3. *Flipping a Coin and Rolling a Die*

Consider the following generative process. An experimenter flips a fair coin. If the coin comes up heads, the experimenter rolls a fair four-sided die. If the coin comes up tails, she rolls a fair six-sided die. The sample space can thus be represented by:

$$\Omega = \{(H,1),(H,2),(H,3),(H,4),(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}.$$

Let A denote the event of observing heads, B denote the event of observing 3, and C denote the event of observing 6. Formally, $A = \{(H,1),(H,2),(H,3),(H,4)\}$, $B = \{(H,3),(T,3)\}$, and $C = \{(T,6)\}$. What is the (joint) probability of observing heads and 3? The probability of observing heads is $P(A) = 1/2$. And if heads is observed, then the experimenter rolls a fair four-sided die, so the probability of observing 3 *given that heads has been observed* is $P(B|A) = 1/4$. So by the Multiplicative Law of Probability,

$$P(A \cap B) = P(B|A)P(A) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}.$$

Likewise, the probability of observing tails and 3 is:

$$P(A^C \cap B) = P(B|A^C)P(A^C) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}.$$

The probability of observing heads and 6 is $P(A \cap C) = P(\emptyset) = 0$.

The conditional probability of observing 3 given that heads (or tails) was observed is straightforward, as we see above. But suppose we wanted to know the conditional probability that heads was observed given that 3 is observed. This is where Bayes' Rule comes in handy. We want to know $P(A|B)$. We know $P(B|A)$ and $P(A)$. What is $P(B)$? From countable additivity,

$$P(B) = P(A \cap B) + P(A^C \cap B) = \frac{1}{8} + \frac{1}{12} = \frac{5}{24}.$$

Thus, by Bayes' Rule,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\frac{1}{4} \cdot \frac{1}{2}}{\frac{5}{24}} = \frac{3}{5}.$$

The "trick" used above to calculate $P(B)$ is actually a special case of another important theorem, the *Law of Total Probability*. To state this theorem, we require the following definition.

Definition 1.1.5. *Partition*

If $A_1, A_2, A_3, ... \in S$ are nonempty and pairwise disjoint, and $\Omega = A_1 \cup A_2 \cup A_3 \cup ...$, then $\{A_1, A_2, A_3, ...\}$ is a partition of Ω.

A partition divides up the sample space into mutually exclusive and exhaustive categories or "bins."[11]

[11] Note that the number of bins may be finite or countably infinite.

Every outcome in Ω is contained in exactly one A_i, so exactly one event A_i in the partition occurs for any draw from (Ω, S, P). We can now state the Law of Total Probability.

Theorem 1.1.5. *Law of Total Probability*

If $\{A_1, A_2, A_3, ...\}$ *is a partition of* Ω *and* $B \in S$*, then*

$$P(B) = \sum_i P(B \cap A_i).$$

If we also have $Pr(A_i) \geq 0$ *for* $i = 1, 2, 3, ...$*, then this can also be stated as*

$$P(B) = \sum_i P(B|A_i)P(A_i).$$

Proof: Let $\{A_1, A_2, A_3, ...\}$ be a partition of Ω and let $B \in S$. Then, $\forall i \neq j$,

$$(B \cap A_i) \cap (B \cap A_j) = (B \cap B) \cap (A_i \cap A_j) = B \cap (A_i \cap A_j) = B \cap \emptyset = \emptyset,$$

so by countable additivity,

$$P\left((B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) \cup ...\right) = \sum_i P(B \cap A_i).$$

And

$$(B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) \cup ... = B \cap (A_1 \cup A_2 \cup A_3 \cup ...) = B \cap \Omega = B,$$

so we have

$$P(B) = \sum_i P(B \cap A_i).$$

Finally, if $P(A_i) > 0$ for $i = 1, 2, 3, ...$, then we can apply the Multiplicative Law of Probability to each term in the summation to obtain

$$P(B) = \sum_i P(B|A_i)P(A_i). \ \square$$

Notice that for any event A, $\{A, A^C\}$ is a partition of Ω, so the method we used to calculate $P(B)$ above was indeed just a special case of the Law of Total Probability. In fact, this "trick" is so often necessary for the application of Bayes' Rule that it is frequently incorporated directly into the statement of the theorem.

Theorem 1.1.6. *Alternative Forms of Bayes' Rule*

If $\{A_1, A_2, A_3, ...\}$ *is a partition of* Ω *with* $P(A_i) > 0$ *for* $i = 1, 2, 3, ...$*, and* $B \in S$ *with* $P(B) > 0$*, then*

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B \cap A_i)},$$

or equivalently,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)}.$$

Proof: Simply apply each form of the Law of Total Probability to the denominator in Theorem 1.1.4. $\square$

1.1.4 Independence of Events

Finally, we define *independence* of events under a random generative process (sometimes referred to as *statistical independence* or *stochastic independence*).

Definition 1.1.6. *Independence of Events*

Events $A, B \in S$ are independent if $P(A \cap B) = P(A)P(B)$.

In Example 1.1.3, events A (observing heads) and B (observing 3) are *not* independent, since $P(A \cap B) = 1/8 \neq 1/2 \cdot 5/24 = P(A)P(B)$. However, if we changed the example so that the experimenter rolled a six-sided die regardless of the outcome of the coin flip, then A and B would be independent.

The following useful theorem follows immediately from this definition and the definition of conditional probability.

Theorem 1.1.7. *Conditional Probability and Independence*

For $A, B \in S$ with $P(B) > 0$, A and B are independent if and only if $P(A|B) = P(A)$.[12]

Proof: Let $A, B \in S$ with $P(B) > 0$. By Definition 1.1.6, A and B are independent $\iff P(A \cap B) = P(A)P(B)$. By Theorem 1.1.3, $P(A|B)P(B) = P(A \cap B)$. Thus, A and B are independent $\iff P(A|B)P(B) = P(A)P(B) \iff P(A|B) = P(A)$. $\square$

Loosely speaking, this theorem tells us that when A and B are independent, knowing whether or not B has occurred gives us *no information* about the probability that A has occurred. Referring again to Example 1.1.3, events A (observing heads) and B (observing 3) are *not* independent, since $P(A|B) = 3/5 \neq 1/2 = P(A)$. The fact that B occurred tells us that it is more likely that A occurred, since overall we expect to observe A (heads) $1/2$ of the time, but among instances in which B (a die roll of 3) is observed, we expect that A will have occurred in $3/5$ of those instances. More extremely, A and C are not independent, since

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(\emptyset)}{P(C)} = 0 \neq \frac{1}{2} = P(A).$$

Overall, the probability of observing heads is $1/2$, but if we know that a 6 was rolled, we can be *certain* that the coin came up tails.

Similarly, $P(B|A) = 1/4 \neq 5/24 = P(B)$. Overall, we expect to observe B (a die roll of 3) $5/24$ of the time, but among the instances in which A is observed (the coin comes up heads), we expect to observe B $1/4$ of the time. So knowing that the coin came up heads increases the probability we assign to observing a die roll of 3. And of course,

$$P(C|A) = \frac{P(C \cap A)}{P(A)} = \frac{P(\emptyset)}{P(A)} = 0 \neq \frac{1}{2} = P(A).$$

[12] Some authors give this as the definition of independence and then prove as a theorem that A and B are independent if and only if $P(A \cap B) = P(A)P(B)$. This formulation is equivalent, and the proof is virtually identical.

When the coin comes up heads, we know for sure that the experimenter will not roll a 6.

Again, if we were to change the example so that the experimenter rolled a six-sided die regardless of the outcome of the coin flip, then A and B would be independent. Knowing the outcome of the coin flip would then tell us nothing about the probability of observing any outcome of the die roll, and vice versa.

1.1.5 Where Are We Going With This?

Randomness makes statistics different from mere description. We can formally describe random generative processes, and *then* data can reveal information about the probabilities involved in the hypothesized processes.

Example 1.1.4. *A Biased Coin Flip*

Consider a coin flip with a (potentially) "biased" coin minted by an unscrupulous character. Let $\Omega = \{H, T\}$ and $S = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$. Then

- $P(\emptyset) = 0$.
- $P(\{H, T\}) = P(\Omega) = 1$.
- $P(\{H\}) = ?$
- $P(\{T\}) = 1 - P(\{H\})$.

We do not have to know $P(\{H\})$ for the generative process to exist. But if we could observe outcomes of this process, we could learn about $P(\{H\})$. If we flipped this biased coin, we would observe heads with probability $P(\{H\})$. By repeating this process, we could learn about $P(\{H\})$. With just one coin flip, we could not make a very good guess about $P(\{H\})$—though we would at least know, if it came up heads, that $P(\{H\}) > 0$, or if it came up tails, that $P(\{H\}) < 1$. With two coin flips, we could make a "better" guess about $P(\{H\})$. With 100 coin flips, we could make an even better guess. What if we could observe infinite coin flips? We could then know the exact value of $P(\{H\})$.

When people talk about *identification*, they are generally speaking about what can be learned about the world given full knowledge of some (Ω, S, P). Variants of the Law of Large Numbers tell us that, given repeated realizations of the same generative process, we can learn everything about S and P (more on this in Chapter 2). We can therefore learn about summaries of S and P, e.g., the expected value (average over repeated draws) given numerical values assigned to outcomes in Ω. Our next task will be describing (Ω, S, P) with something that's a little easier to work with, given numerical values for events.

1.2 Random Variables

From this point on, we will mainly be concerned with random generative processes in which outcomes can be assigned real number values. A *random variable* is a variable that takes on a value that is determined

by such a generative process. For example, in the case of a fair die roll, we might let the random variable X take on the value of the outcome of the die roll. Formally, the value of a random variable is given by a real-valued function of the outcome of a random generative process.

Definition 1.2.1. *Random Variable*

A random variable X is a variable whose value is given by $X = \mathcal{X}(\omega)$, where $\mathcal{X} : \Omega \to \mathbb{R}$ such that, $\forall r \in \mathbb{R}$, $\{\omega \in \Omega : \mathcal{X}(\omega) \leq r\} \in S$.[13]

Recall that each $\omega \in \Omega$ denotes a state of the world, which may be represented by anything: numbers, letters, words, or other symbols, whatever notation is most convenient to describe all of the distinct possible outcomes that could occur. $\mathcal{X}(\omega)$ then tells us the value that X takes on when the state of the world is ω. In other words, just as we write $y = f(x)$ to denote that the value of y is determined by applying the "rule" f to the value of x, we write $X = \mathcal{X}(\omega)$ to denote that the value of the random variable X is determined by applying the "rule" $\mathcal{X}$ to the "value" of ω.

We can define events in S in terms of a random variable X. For example, we could let

- $A = \{\omega \in \Omega : \mathcal{X}(\omega) = 1\} = \{\omega \in \Omega : X = 1\}$.
- $B = \{\omega \in \Omega : \mathcal{X}(\omega) \geq 0\} = \{\omega \in \Omega : X \geq 0\}$.
- $C = \{\omega \in \Omega : \mathcal{X}(\omega)^2 < 10, \mathcal{X}(\omega) \neq 3\} = \{\omega \in \Omega : X^2 < 10, X \neq 3\}$.[14]

As shorthand, we typically drop the $\omega \in \Omega$ and simply denote events by their defining conditions on X. So, for the above examples, we would write

- $A = \{X = 1\}$.
- $B = \{X \geq 0\}$.
- $C = \{X^2 < 10, X \neq 3\}$.

Furthermore, we use $\Pr(\cdot)$ as shorthand to denote the probability that some condition(s) on a random variable[15] will hold, or, equivalently, the probability that the event defined by those conditions will occur. So, for the above examples, we would write

- $\Pr(X = 1) = P(A)$.
- $\Pr(X \geq 0) = P(B)$.
- $\Pr(X^2 < 10, X \neq 3) = P(C)$.

[13] This regularity condition is necessary and sufficient to ensure that $\mathcal{X}$ is a measurable function. In particular, it ensures that we can define the cumulative distribution function of X (see Definition 1.2.3).

[14] Note that, when we list conditions separated by commas, the commas implicitly mean "and."

[15] Or variables; see Section 1.3.

Note that we use uppercase letters $(X, Y, Z, W, ...)$ to denote random variables, and cursive uppercase letters $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}, ...)$ to denote the functions that determine the values of random variables. We use lowercase letters $(x, y, z, w, ...)$ as "placeholders" for specific outcomes[16] of random variables, i.e., as variables in the regular, algebraic sense. (See, e.g., Definition 1.2.2.)

1.2.1 Discrete Random Variables

A *discrete random variable* is a random variable that can take on a finite or countably infinite number of different values. That is, a random variable X is discrete if the range of $\mathcal{X}$, $\mathcal{X}(\Omega)$, is a countable set.

Example 1.2.1. *A Fair Die Roll*

Consider again a roll of one fair (six-sided) die. Let X take on the value of the outcome of the die roll, i.e., let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $X = \mathcal{X}(\omega) = \omega$, $\forall \omega \in \Omega$. Then $\Pr(X = 1) = \Pr(X = 2) = ... = \Pr(X = 6) = 1/6$. Out of many die rolls, we expect each of the values 1 through 6 to come up $1/6$ of the time.

Example 1.2.2. *A Biased Coin Flip*

Consider, again, a coin flip with a (potentially) biased coin. Let $X = 0$ if the coin comes up heads and $X = 1$ if the coin comes up tails, i.e., let $\Omega = \{H, T\}$ and let $\mathcal{X}(H) = 0$ and $\mathcal{X}(T) = 1$. Let p be the probability that the coin comes up tails: $\Pr(X = 1) = p$. Then $\Pr(X = 0) = 1 - \Pr(X = 1) = 1 - p$. Out of many coin flips, we expect that the proportion of times the coin comes up tails will be p and the proportion of times the coin comes up heads will be $1 - p$.

A discrete random variable X that takes on the value 0 or 1 with $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$ (where $0 < p < 1$) is called a *Bernoulli random variable* or *binary random variable*. This type of random variable will be recurrent in later chapters.

Given a discrete random variable X, we can summarize the probability of each outcome x occurring with a *probability mass function (PMF)*.

Definition 1.2.2. *Probability Mass Function (PMF)*

For a discrete random variable X, the probability mass function of X is:

$$f(x) = \Pr(X = x), \forall x \in \mathbb{R}.^{17}$$

In the above example of a fair die roll, the PMF of X is:

$$f(x) = \begin{cases} \frac{1}{6} & : x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & : \text{otherwise} \end{cases}$$

[16] We will informally refer to elements $x \in \mathcal{X}(\Omega)$ as outcomes, even though, strictly speaking, they are not necessarily elements of the sample space Ω, and, indeed, there may be more than one $\omega \in \Omega$ such that $\mathcal{X}(\omega) = x$.

[17] Alternatively, $f(x)$ may be left undefined for $x \notin \mathcal{X}(\Omega)$. The distinction is unimportant for our purposes.

For a Bernoulli random variable (e.g., the biased coin flip), the PMF of X is:

$$f(x) = \begin{cases} 1-p & : x = 0 \\ p & : x = 1 \\ 0 & : \text{otherwise} \end{cases}$$

For a discrete random variable, the PMF tells us *everything* about its distribution, i.e., we can give the probability of *any* event that can be defined just in terms of X. For example, consider a discrete random variable X such that $f(x) = 0$ for all $x \notin \mathbb{Z}$ (i.e., X takes on only integer values). Then

- $\Pr(X \geq 3) = \sum_{x \in \{3,4,5,\ldots\}} f(x)$.
- $\Pr(X \geq 3 \text{ or } X = 1) = \sum_{x \in \{1,3,4,5,\ldots\}} f(x)$.
- $\Pr(X \geq 3, X = 1) = 0$.

E.g., for the die roll (Example 1.2.1),

- $\Pr(X \geq 3) = \sum_{x \in \{3,4,5,6\}} f(x) = \frac{4}{6} = \frac{2}{3}$.
- $\Pr(X \geq 3 \text{ or } X = 1) = \sum_{x \in \{1,3,4,5,6\}} f(x) = \frac{5}{6}$.
- $\Pr(X \geq 3, X = 1) = 0$.

More generally, the following theorem gives the formula for using the PMF to compute the probability of *any* event defined in terms of a discrete random variable X.

Theorem 1.2.1. *Event Probabilities for Discrete Random Variables*

For a discrete random variable X with PMF f, if $D \subseteq \mathbb{R}$ and $A = \{X \in D\}$, then

$$\Pr(X \in D) = \sum_{x \in \mathcal{X}(A)} f(x).$$

We omit the proof of this theorem. Note that any condition on X can be expressed as $X \in D$ for some set $D \subseteq \mathbb{R}$, so this theorem allows us to compute the probability of any event defined in terms of a discrete random variable X. We now define an alternative (and more general) way of describing the distribution of a random variable, the *cumulative distribution function* (CDF).

Definition 1.2.3. *Cumulative Distribution Function (CDF)*

For a random variable X, the cumulative distribution function of X is:

$$F(x) = \Pr(X \leq x), \forall x \in \mathbb{R}.$$

Quite simply, the CDF returns the probability that an outcome for a random variable will be less than or equal to a given value. Importantly, given any random variable X, the CDF of X tells us *everything* there is to know about X. For any event A that can be described just in terms of X, we can derive the probability of A from the CDF of X alone. The following important properties of CDFs follow immediately from the axioms and basic properties of probability.

Theorem 1.2.2. *Properties of CDFs*

For a random variable X with CDF F,

- *F is nondecreasing: $\forall x_1, x_2 \in \mathbb{R}$, if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$.*
- $\lim_{x \to -\infty} F(x) = 0$.
- $\lim_{x \to \infty} F(x) = 1$.
- $\forall x \in \mathbb{R}$, $1 - F(x) = \Pr(X > x)$.

Proof: Let X be a random variable with CDF F. Let $x_1, x_2 \in \mathbb{R}$ with $x_1 < x_2$. Then since $\{X \leq x_1\} \subseteq \{X \leq x_2\}$, monotonicity implies F is nondecreasing:

$$F(x_1) = \Pr(X \leq x_1) = P(\{X \leq x_1\}) \leq P(\{X \leq x_2\}) = \Pr(X \leq x_2) = F(x_2).$$

P is continuous,[18] and $\lim_{x \to -\infty} \{X \leq x\} = \{X \in \emptyset\} = \emptyset$, so

$$\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} \Pr(X \leq x) = \lim_{x \to -\infty} P(\{X \leq x\}) = P\left(\lim_{x \to -\infty} \{X \leq x\}\right) = P(\emptyset) = 0.$$

Similarly, $\lim_{x \to \infty} \{X \leq x\} = \{X \in \mathbb{R}\} = \Omega$, so

$$\lim_{x \to \infty} F(x) = \lim_{x \to \infty} \Pr(X \leq x) = \lim_{x \to \infty} P\left(\{X \leq x\}\right) = P\left(\lim_{x \to \infty} \{X \leq x\}\right) = P(\Omega) = 1.$$

Finally, by the complement rule, $\forall x \in \mathbb{R}$,

$$1 - F(x) = 1 - \Pr(X \leq x) = 1 - P(\{X \leq x\}) = P(\{X \leq x\}^C) = P(\{X > x\}) = \Pr(X > x). \ \square$$

[18] More precisely, for any sequence of sets $A_1, A_2, A_3, ... \in S$, if $A_1 \subseteq A_2 \subseteq A_3 \subseteq ...$ and $A_1 \cup A_2 \cup A_3 \cup ... = A$, then

$$\lim_{i \to \infty} P(A_1 \cup A_2 \cup ... \cup A_i) = P(A),$$

and for any sequence of sets $B_1, B_2, B_3, ... \in S$, if $B_1 \supseteq B_2 \supseteq B_3 \supseteq ...$ and $B_1 \cap B_2 \cap B_3 \cap ... = B$, then

$$\lim_{i \to \infty} P(B_1 \cap B_2 \cap ... \cap B_i) = P(B).$$

This property allows us to "pass" the limit into $P(\,\cdot\,)$. We omit the proof of this fact.

Returning again to the die roll example:

- $F(1) = \frac{1}{6}$
- $F(1.5) = \frac{1}{6}$
- $F(2) = \frac{2}{6} = \frac{1}{3}$
- $F(6) = 1$
- $F(24603) = 1$
- $\Pr(X \geq 3) = 1 - \Pr(X < 3) = 1 - \Pr(X \leq 2) = 1 - F(2) = \frac{4}{6} = \frac{2}{3}$
- $\Pr(X < 2) = \Pr(X \leq 1) = F(1) = \frac{1}{6}$
- $\Pr(2 \leq X \leq 4) = \Pr(X \leq 4) - \Pr(X < 2) = F(4) - F(1) = \frac{4}{6} - \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$,

where the first step of the last line follows from the subtraction rule (see Theorem 1.1.1).
For discrete random variables, the CDF is a step function. Figure 1.2.1 shows the CDF for the die roll.

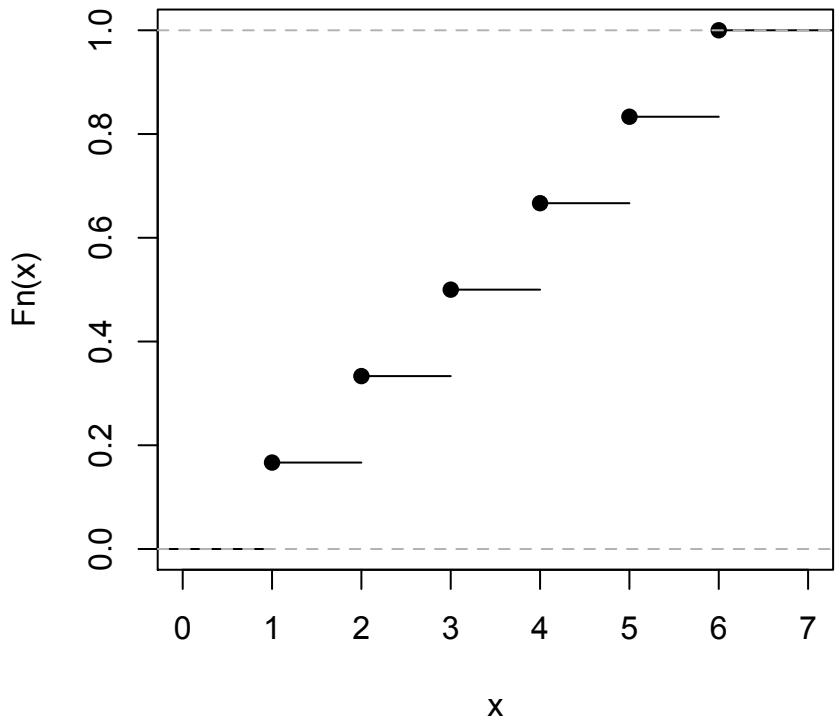


Figure 1.2.1 *Die Roll CDF*

1.2.2 Continuous Random Variables

Loosely speaking, a *continuous random variable* is a random variable that has a continuous CDF.[19] This implies that a continuous random variable X can take on an uncountably infinite number of different values, i.e., $\mathcal{X}(\Omega)$ is an uncountable set (and therefore so is Ω). Typically $\mathcal{X}(\Omega)$ will be some interval or union of intervals of the real line.

Definition 1.2.4. *Continuous Random Variable*

A random variable X is continuous if there exists a nonnegative function $f : \mathbb{R} \to \mathbb{R}$ such that the CDF of X is:

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(t)dt, \forall x \in \mathbb{R}.$$

The function f is called the probability density function (PDF).

The following theorem defines the PDF more explicitly.

Theorem 1.2.3. *Probability Density Function (PDF)*

For a continuous random variable X with CDF F, the probability density function of X is:

$$f(x) = \left.\frac{dF(t)}{dt}\right|_{t=x}, \forall x \in \mathbb{R}.$$

Proof: This follows directly from the Fundamental Theorem of Calculus. □

Conceptually, the PDF is the continuous analog to the PMF in that it describes how the CDF changes with x. The difference is that, whereas a PMF specifies the size of the "jump" in the CDF at a point x, a PDF gives the instantaneous slope (derivative) of the CDF at a point x. That is, for a very small number $\epsilon > 0$, if we moved from x to $x + \epsilon$, the CDF would change by approximately $\epsilon f(x)$.

We now note some important properties of PDFs.

Theorem 1.2.4. *Properties of PDFs*

For a continuous random variable X with PDF f,

- $\forall x \in \mathbb{R}, f(x) \geq 0$.
- $\int_{-\infty}^{\infty} f(x)dx = 1$.

[19] Technically, this is a necessary but not sufficient condition for a random variable to be continuous. The sufficient condition is that the CDF be absolutely continuous with respect to Lebesgue measure. This is a technical measure-theoretic condition that need not concern us, as it is implied by Definition 1.2.4.

Proof: Let X be a continuous random variable with CDF F and PDF f. F is nondecreasing, so, $\forall x \in \mathbb{R}$,

$$f(x) = \left.\frac{dF(t)}{dt}\right|_{t=x} \geq 0.$$

And

$$\int_{-\infty}^{\infty} f(x)dx = \lim_{x\to\infty} \int_{-\infty}^{x} f(t)dt = \lim_{x\to\infty} F(x) = 1. \ \square$$

The following theorem gives the formula for using the PDF to compute the probability of any event of the form $\{X \in I\}$ where I is an interval in $\mathbb{R}$.

Theorem 1.2.5. *Event Probabilities for Continuous Random Variables*

For a continuous random variable X with PDF f,

- $\forall x \in \mathbb{R}, \Pr(X = x) = 0$.
- $\forall x \in \mathbb{R}, \Pr(X < x) = \Pr(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)dt$.
- $\forall x \in \mathbb{R}, \Pr(X > x) = \Pr(X \geq x) = 1 - F(x) = \int_{x}^{\infty} f(t)dt$.
- $\forall a, b \in \mathbb{R}$ *with* $a \leq b, \Pr(a < X < b) = \Pr(a \leq X < b) = \Pr(a < X \leq b) = \Pr(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$.

Proof: Let X be a continuous random variable with CDF F and PDF f. Let $x \in \mathbb{R}$. To establish that, $\forall x \in \mathbb{R}, \Pr(X = x) = 0$, we will proceed with a proof by contradiction. Suppose that $\Pr(X = x) = P(\{X = x\}) > 0$. Then since P is continuous and $\lim_{t\to x^-}\{X \leq t\} = \{X < x\}$,

$$\begin{aligned}
\lim_{t\to x^-} F(t) &= \lim_{t\to x^-} \Pr(X \leq t) \\
&= \lim_{t\to x^-} P(\{X \leq t\}) \\
&= P\left(\lim_{t\to x^-} \{X \leq t\}\right) \\
&= P(\{X < x\}) \\
&< P(\{X < x\}) + P(\{X = x\}) \\
&= P(\{X \leq x\}) \\
&= \Pr(X \leq x) \\
&= F(x),
\end{aligned}$$

Thus, F has a discontinuity at x, contradicting the assumption that X is continuous. So $\Pr(X = x) \leq 0$,

and thus, by nonegativity of probability, $\Pr(X = x) = 0$. Then $\forall x \in \mathbb{R}$,

$$\begin{aligned}
\Pr(X < x) &= P(\{X < x\}) \\
&= P(\{X < x\}) + P(\{X = x\}) \\
&= P(\{X \leq x\}) \\
&= \Pr(X \leq x) \\
&= F(x) \\
&= \int_{-\infty}^{x} f(t)dt.
\end{aligned}$$

Similarly, $\forall x \in \mathbb{R}$,

$$\begin{aligned}
\Pr(X \geq x) &= P(\{X \geq x\}) \\
&= P(\{X > x\}) + P(\{X = x\}) \\
&= P(\{X > x\}) \\
&= \Pr(X > x) \\
&= 1 - F(x) \\
&= 1 - \int_{-\infty}^{x} f(t)dt \\
&= \int_{x}^{-\infty} f(t)dt + 1 \\
&= \int_{x}^{-\infty} f(t)dt + \int_{-\infty}^{\infty} f(t)dt \\
&= \int_{x}^{\infty} f(t)dt.
\end{aligned}$$

Finally, by the same logic as above, $\forall a, b \in \mathbb{R}$ with $a < b$,

$$\Pr(a < X < b) = \Pr(a \leq X < b) = \Pr(a < X \leq b) = \Pr(a \leq X \leq b).$$

Furthermore,

$$P(\{X < a\}) + P(\{a \leq X \leq b\}) = P(\{X \leq b\}),$$

so

$$P(\{a \leq X \leq b\}) = P(\{X \leq b\}) - P(\{X < a\}),$$

and thus

$$\begin{aligned}
\Pr(a \leq X \leq b) &= P(\{a \leq X \leq b\}) \\
&= P(\{X \leq b\}) - P(\{X < a\}) \\
&= \Pr(X \leq b) - \Pr(X < a) \\
&= F(b) - F(a) \\
&= \int_{a}^{b} f(x)dx,
\end{aligned}$$

where the final equality follows from the Fundamental Theorem of Calculus. $\square$

Note that by applying countable additivity, we can use Theorem 1.2.5 to compute the probability of any event of the form $\{X \in D\}$, where D is a countable union of disjoint intervals in $\mathbb{R}$.[20]

It may seem strange that, for a continuous random variable X, any specific outcome $x \in \mathbb{R}$ has probability $\Pr(X = x) = 0$, but this should not bother us too much. Suppose we chose a random person from some population and measured his or her weight. If our scale were extremely precise, then it would be very unlikely that we would observe exactly, say, 147.627408 lbs. The greater the precision of the scale, the less likely it would be that we would observe any specific value. With *infinite* precision, the probability of observing any exact weight would be 0. In the real world, everything is discrete, because we never have infinite precision of measurement, but continuous distributions are a mathematically convenient way of modeling random generative processes when outcomes are very fine-grained.

Theorem 1.2.5 provides us with a clearer intuition for the meaning of the PDF: the area under the PDF over an interval is equal to the probability that the random variable takes on a value in that interval. Note that, formally, this theorem implies that we need not worry about strict versus weak inequalities when describing the probability that a random variable falls within a certain interval.

The set of values at which the PMF or PDF of a random variable is positive is called its *support*.

Definition 1.2.5. *Support*

For a random variable X with PMF/PDF f, the support of X is

$$\text{Supp}\,(X) = \mathcal{X}(\Omega) = \{x \in \mathbb{R} \,:\, f(x) > 0\}.$$

For discrete X, $\text{Supp}\,(X)$ is the set of values that X takes on with positive probability. For continuous X, $\text{Supp}\,(X)$ is the set of values over which X has positive probability density.

We now discuss two important examples of continuous distributions, the *standard uniform distribution* and the *standard normal distribution*.

Example 1.2.3. *The Standard Uniform Distribution*

Consider the standard uniform distribution, denoted $U(0, 1)$. Informally, if a random variable X follows the standard uniform distribution, X takes on a random real value from the interval $[0, 1]$, with all values in this interval equally likely to occur. We write $X \sim U(0, 1)$ to denote that X follows the standard uniform distribution. The standard uniform distribution has the PDF

$$f(x) = \begin{cases} 1 & : 0 \leq x \leq 1 \\ 0 & : \text{otherwise} \end{cases}$$

Figure 1.2.2 plots the PDF of the standard uniform distribution.

[20] In measure theory terms, for any Lebesgue-measurable set $D \subseteq \mathbb{R}$, $\Pr(X \in D) = \int_D f(x)dx$.

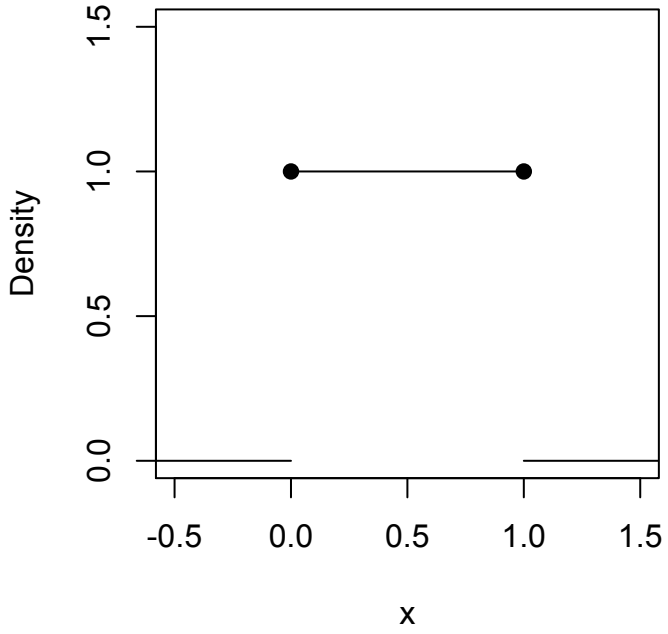


Figure 1.2.2 *PDF of Standard Uniform Distribution*

The CDF of $U(0,1)$ is thus

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 & : x < 0 \\ x & : 0 \leq x \leq 1 \\ 1 & : x > 1 \end{cases}$$

Figure 1.2.3 shows the CDF of the standard uniform distribution.

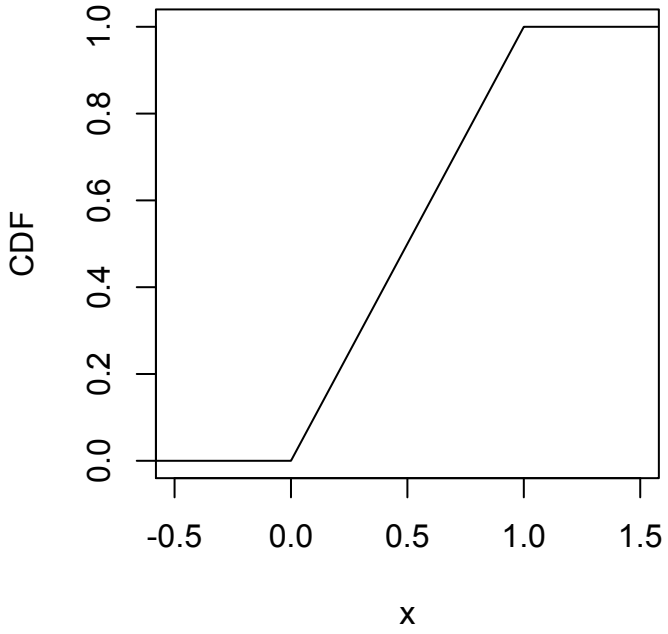


Figure 1.2.3 *CDF of Standard Uniform Distribution*

Example 1.2.4. *The Standard Normal Distribution*

You are probably familiar with the standard normal (or Gaussian) distribution. There are many reasons why the standard normal distribution is important, some of which will be discussed later in this book (see Section 2.1.4). The standard normal distribution, denoted $N(0, 1)$, has PDF

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

and CDF

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Note that, by convention, $\phi(x)$ and $\Phi(x)$ denote the PDF and CDF of the standard normal distribution, respectively. These are shown in Figures 1.2.4 and 1.2.5.

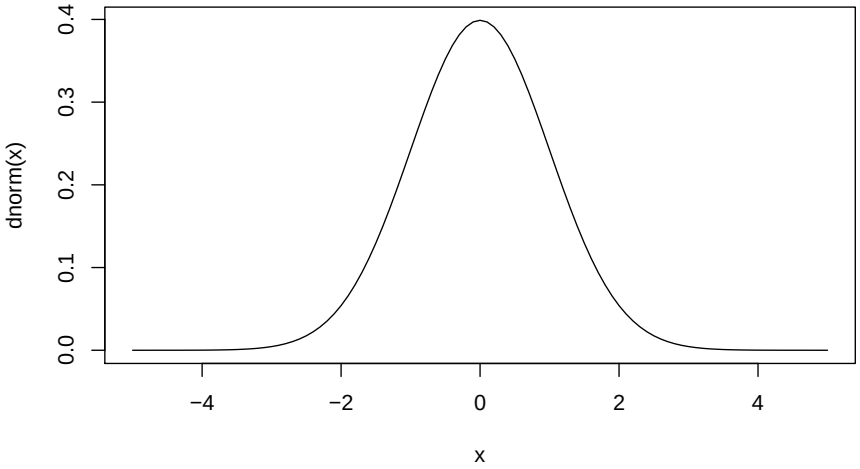


Figure 1.2.4 *PDF of Standard Normal Distribution*

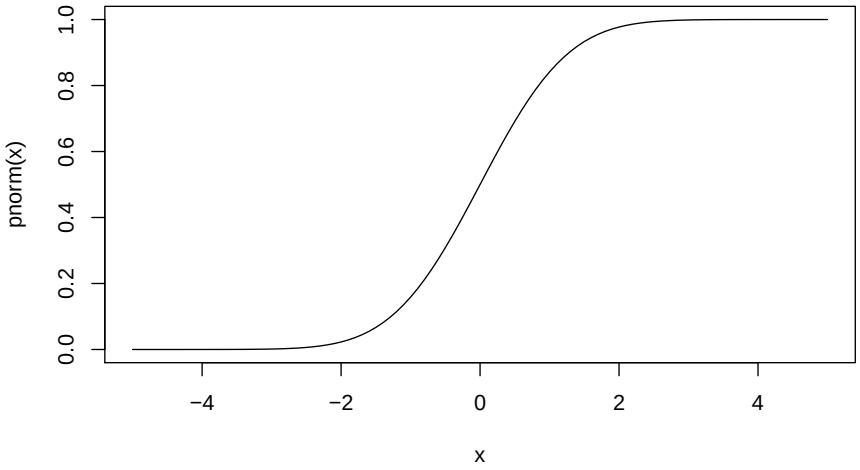


Figure 1.2.5 *CDF of Standard Normal Distribution*

1.3 Bivariate Relationships

As with random events, we often want to describe how multiple random variables are related. We therefore generalize the concepts presented above to the bivariate case.[21] A *random vector* (X, Y) is a pair of associated random variables whose values are determined simultaneously by a single generative process, i.e., $(X, Y) = (\mathcal{X}(\omega), \mathcal{Y}(\omega))$, where $\mathcal{X} : \Omega \to \mathbb{R}$ and $\mathcal{Y} : \Omega \to \mathbb{R}$, with $\mathcal{X}$ and $\mathcal{Y}$ each satisfying the regularity condition in Definition 1.2.1.

[21] Further generalization to the case of three or more random variables can be done analogously (see Section 1.4.9).

The following fundamental definition will be important later: when we say two random variables are equal, we mean that they take on the same value in every state of the world.

Definition 1.3.1. *Equality of Random Variables*

Let X and Y be random variables whose values are given by $X = \mathcal{X}(\omega)$ and $Y = \mathcal{Y}(\omega)$. We say that $X = Y$ if, $\forall \omega \in \Omega$,

$$\mathcal{X}(\omega) = \mathcal{Y}(\omega).$$

This is also equivalent to saying that $\Pr(X = Y) = 1$.

1.3.1 Bivariate Distributions

In the case where X and Y are both discrete random variables, we can define the *joint PMF* of X and Y. The joint PMF at (x, y) is simply the probability that $X = x$ *and* $Y = y$ in a single realization of (X, Y).

Definition 1.3.2. *Joint PMF*

For discrete random variables X and Y, the joint PMF of X and Y is:

$$f(x, y) = \Pr(X = x, Y = y), \forall (x, y) \in \mathbb{R}^2.$$

Likewise, for any random variables X and Y, we can define their *joint CDF*. The joint CDF at (x, y) gives the probability of observing $X \leq x$ *and* $Y \leq y$ in a single realization of (X, Y).

Definition 1.3.3. *Joint CDF*

For random variables X and Y, the joint CDF of X and Y is:

$$F(x, y) = \Pr(X \leq x, Y \leq y), \forall (x, y) \in \mathbb{R}^2.$$

Example 1.3.1. *Flipping a Coin and Rolling a Die*

Consider the generative process from Example 1.1.3, in which the experimenter flips a coin and then rolls either a four-sided or six-sided die depending on the outcome of the coin flip. Let $X = 0$ if the coin comes up heads and $X = 1$ if the coin comes up tails, and let Y be the value of the outcome of the die roll. Then the joint PMF of (X, Y) is

$$f(x, y) = \begin{cases} \frac{1}{8} & : x = 0, y \in \{1, 2, 3, 4\} \\ \frac{1}{12} & : x = 1, y \in \{1, 2, 3, 4, 5, 6\} \\ 0 & : \text{otherwise} \end{cases}$$

For discrete random variables, the joint CDF is constructed simply by summing over the appropriate values

of X and Y. So in this case, for example,

$$\begin{aligned} F(1,3) &= \sum_{x \leq 1} \sum_{y \leq 3} f(x,y) \\ &= f(0,1) + f(0,2) + f(0,3) + f(1,1) + f(1,2) + f(1,3) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} = \frac{5}{8}. \end{aligned}$$

In addition to the joint PMF and joint CDF, we can also describe the distribution of (X, Y) in terms of the conditional and marginal PMFs of X and Y. The *marginal PMF* of X is simply the PMF of X, ignoring the existence of Y.

Definition 1.3.4. *Marginal PMF*

For discrete random variables X and Y with joint PMF f, the marginal PMF of X is:

$$f_X(x) = \Pr(X = x) = \sum_{y \in \text{Supp}\,(Y)} f(x,y), \forall x \in \mathbb{R}.$$

Likewise, the marginal PMF of Y is

$$f_Y(y) = \Pr(Y = y) = \sum_{x \in \text{Supp}\,(X)} f(x,y), \forall y \in \mathbb{R}.^{22}$$

Note the relationship here to the Law of Total Probability (Theorem 1.1.5): we sum up the joint probabilities of $X = x$ and $Y = y$ for every possible outcome for Y to get the overall probability that $X = x$.

The *conditional PMF* of X given Y tells us the probability that a given value of X will occur, *given that* a certain value of Y occurs. That is, the conditional PMF of X given $Y = y$ tells us what the PMF of X would be if we "threw out" all realizations of (X, Y) except those in which $Y = y$.

Definition 1.3.5. *Conditional PMF*

For discrete random variables X and Y with joint PMF f, the conditional PMF of X given $Y = y$ is:

$$f_{X|Y}(x|y) = \Pr(X = x|Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} = \frac{f(x,y)}{f_Y(y)}, \forall (x,y) \in \mathbb{R}^2 \textit{ with } f_Y(y) > 0.$$

Likewise, the conditional PMF of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \Pr(Y = y|X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} = \frac{f(x,y)}{f_X(x)}, \forall (x,y) \in \mathbb{R}^2 \textit{ with } f_X(x) > 0.$$

This is simply the definition of conditional probability applied to PMFs.

[22] Henceforth we shall abbreviate $\sum_{x \in \text{Supp}\,(X)}$ as $\sum_x$.

Example 1.3.2. *Flipping a Coin and Rolling a Die*

In the coin flip and die roll example, the marginal PMFs are:

$$f_X(x) = \begin{cases} \frac{1}{2} & : x = 0 \\ \frac{1}{2} & : x = 1 \\ 0 & : \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{5}{24} & : y \in \{1, 2, 3, 4\} \\ \frac{1}{12} & : y \in \{5, 6\} \\ 0 & : \text{otherwise} \end{cases}$$

and the conditional PMFs are:

$$f_{X|Y}(x|y) = \begin{cases} \frac{3}{5} & : x = 0, y \in \{1, 2, 3, 4\} \\ \frac{2}{5} & : x = 1, y \in \{1, 2, 3, 4\} \\ 1 & : x = 1, y \in \{5, 6\} \\ 0 & : \text{otherwise} \end{cases}$$

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{4} & : x = 0, y \in \{1, 2, 3, 4\} \\ \frac{1}{6} & : x = 1, y \in \{1, 2, 3, 4, 5, 6\} \\ 0 & : \text{otherwise} \end{cases}$$

Note that we can operate on marginal PMFs in the same way that we would univariate PMFs, since they *are* univariate PMFs. Similarly, conditional PMFs are univariate PMFs given any fixed value for the conditioning variable. E.g., in this example,

$$f_{X|Y}(x|3) = \begin{cases} \frac{3}{5} & : x = 0 \\ \frac{2}{5} & : x = 1 \\ 0 & : \text{otherwise} \end{cases}$$

is a univariate PMF.

What about continuous distributions? Conceptually, everything is the same; generalizing continuous distributions to the bivariate case proceeds analogously to generalizing discrete distributions.

Definition 1.3.6. *Jointly Continuous Random Variables*

Two random variables X and Y are jointly continuous if there exists a nonnegative function $f : \mathbb{R}^2 \to \mathbb{R}$ such that the joint CDF of X and Y is:

$$F(x, y) = \Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) dv du, \forall (x, y) \in \mathbb{R}^2.$$

The function f is called the joint probability density function (joint PDF).

Just as a single continuous random variable is characterized by a continuous CDF, two jointly continuous random variables are characterized by a continuous joint CDF. And just as taking the derivative of the CDF of a continuous random variable yields the PDF, taking the mixed second-order partial derivative of the joint CDF of two jointly continuous random variables yields the joint PDF.

Definition 1.3.7. *Joint PDF*

For jointly continuous random variables X and Y with joint CDF F, the joint PDF of X and Y is:

$$f(x, y) = \left. \frac{\partial^2 F(u, v)}{\partial u \partial v} \right|_{u=x, v=y}, \forall (x, y) \in \mathbb{R}^2.$$

Without going into too much detail regarding partial derivatives, the joint PDF has roughly the same interpretation as the univariate PDF: if we perturb x or y a little bit, how much does the CDF change? Perhaps more intuitively, as with univariate continuous distributions, event probabilities are computed by integrating: $\forall a, b, c, d \in \mathbb{R}$ with $a \leq b$ and $c \leq d$,

$$\Pr(a \leq X \leq b, c \leq Y \leq d) = \int_{a}^{b} \int_{c}^{d} f(x, y) dy dx.$$

i.e., the volume under the PDF over a region equals the probability that the random vector (X, Y) takes on a value in that region. Indeed, the probability of *any* event (not just those represented by rectangular regions) can be computed by integration.

Theorem 1.3.1. *Event Probabilities for Bivariate Continuous Distributions*

For jointly continuous random variables X and Y with joint PDF f, if $D \subseteq \mathbb{R}^2$, then

$$\Pr\Big((X, Y) \in D\Big) = \iint_{D} f(x, y) dy dx.^{23}$$

We omit the proof of this theorem. So, for example, the probability that (X, Y) will fall within the triangular region $\{0 \leq X \leq 1, Y \leq X\}$ is:

$$\Pr(0 \leq X \leq 1, Y \leq X) = \int_{0}^{1} \int_{0}^{x} f(x, y) dy dx.$$

[23] As in the univariate case, D must of course be Lebesgue-measurable, or the integral does not exist.

Consequently, properties analogous to those in Theorem 1.2.4 and Theorem 1.2.5 apply to bivariate continuous distributions: the joint PDF is nonnegative everywhere and integrates to 1; exact outcomes have probability zero,[24] and therefore strict versus weak inequalities in event specifications make no difference, e.g., $\Pr(X \geq 7, 3 \leq Y < 6) = \Pr(X > 7, 3 < Y \leq 6)$ if X and Y are jointly continuous.

We can now define marginal and conditional PDFs analogously to marginal and conditional PMFs. As in the discrete case, the *marginal PDF* of X is simply the PDF of X, ignoring the existence of Y, while the *conditional PDF* of X given Y is the PDF of X *given that* a certain value of Y occurs.

Definition 1.3.8. *Marginal PDF*

For jointly continuous random variables X and Y with joint PDF f, the marginal PDF of X is:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy, \forall x \in \mathbb{R}.$$

Likewise, the marginal PDF of Y is:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx, \forall y \in \mathbb{R}.$$

Definition 1.3.9. *Conditional PDF*

For jointly continuous random variables X and Y with joint PDF f, the conditional PDF of X given $Y = y$ is:

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}, \forall (x, y) \in \mathbb{R}^2 \text{ with } f_Y(y) > 0.$$

Likewise, the conditional PDF of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}, \forall (x, y) \in \mathbb{R}^2 \text{ with } f_X(x) > 0.$$

As in the discrete case, marginal PDFs are univariate PDFs, and conditional PDFs are univariate PDFs given any fixed value for the conditioning variable. Theorem 1.2.4 thus applies to all marginal PDFs and to all conditional PDFs given any fixed value of the conditioning variable.

Notice that we can rearrange Definition 1.3.5 or Definition 1.3.9 to obtain an analog to the Multiplicative Law of Probability (Theorem 1.1.3) for PMFs/PDFs.

Theorem 1.3.2. *Multiplicative Law for PMFs/PDFs*

Let X and Y be either two discrete random variables with joint PMF $f(x, y)$ or two jointly continuous random variables with joint PDF $f(x, y)$. Then $\forall (x, y) \in \mathbb{R}^2$ with $f_Y(y) > 0$,

$$f_{X|Y}(x|y)f_Y(y) = f(x, y).$$

[24] Note that, for bivariate continuous distributions, this means not only that $\Pr(X = x, Y = y) = 0$, but also that any event that specifies *either* $X = x$ *or* $Y = y$ has probability zero, e.g., $\Pr(X > x, Y = y) = 0$. Indeed, for any $D \subseteq \mathbb{R}^2$ that has zero area (i.e., Lebesgue measure zero), $\Pr((X, Y) \in D) = 0$, e.g., $\Pr(X = Y) = 0$, since the set $\{(x, y) \in \mathbb{R}^2 \mid x = y\}$ is a line.

Proof: Simply rearrange Definition 1.3.5 (discrete case) or Definition 1.3.9 (continuous case). □

Similarly, we can derive an analog to Bayes' Rule for PMFs/PDFs.

Theorem 1.3.3. *Bayes' Rule for PMFs/PDFs*

Let X and Y be either two discrete random variables with joint PMF $f(x, y)$ or two jointly continuous random variables with joint PDF $f(x, y)$. Then, $\forall (x, y) \in \mathbb{R}^2$ with $f_X(x) > 0$ and $f_Y(y) > 0$,

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}.$$

Proof: In the discrete case, the proof is immediate: just apply Bayes' Rule with $A = \{X = x\}$ and $B = \{Y = y\}$. In the continuous case, we proceed in the same fashion as the proof of Bayes' Rule. Let $(x, y) \in \mathbb{R}^2$ with $f_X(x) > 0$ and $f_Y(y) > 0$. By the Multiplicative Law for PDFs,

$$f(x, y) = f_{Y|X}(y|x) f_X(x).$$

So by the definition of conditional PDF,

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}. \ \square$$

The reader can verify that Theorem 1.3.2 and Theorem 1.3.3 hold for the coin flip and die roll example. One can also use the definition of marginal PMF/PDF (recalling that this is analogous to the Law of Total Probability) and the Multiplicative Law for PMFs/PDFs to obtain alternative forms of Bayes' Rule for PMFs/PDFs, analogous to the alternative forms of Bayes' Rule given by Theorem 1.1.6. We leave this as an exercise for the reader.

1.3.2 Independence of Random Variables

Like random events, random variables can be (statistically/stochastically) independent. We conclude this section by defining independence of random variables and noting some important implications of independence.

Definition 1.3.10. *Independence of Random Variables*

Let X and Y be either two discrete random variables with joint PMF $f(x, y)$ or two jointly continuous random variables with joint PDF $f(x, y)$. Then X and Y are independent if, $\forall (x, y) \in \mathbb{R}^2$,

$$f(x, y) = f_X(x) f_Y(y).$$

We write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent.

Note the similarity to the definition of independence of events (Definition 1.1.6). Returning again to the coin flip and die roll example, we see that X (the outcome of the coin flip, where $X = 0$ for heads and $X = 1$ for tails) and Y (the outcome of the die roll) clearly are *not* independent, since, as we have already shown, $f_{X|Y}(0|3) = 3/5 \neq 1/2 \cdot 5/24 = f_X(0)f_Y(3)$.

The following theorem states some properties that hold if and only if X and Y are independent.

Theorem 1.3.4. *Implications of Independence (Part I)*

Let X and Y be either two discrete random variables with joint PMF $f(x, y)$ or two jointly continuous random variables with joint PDF $f(x, y)$, and let $F(x, y)$ be the joint CDF of X and Y. Then the following statements are equivalent (i.e., each one implies all the others):

- $X \perp\!\!\!\perp Y$.
- $\forall (x, y) \in \mathbb{R}^2, F(x, y) = F(x)F(y)$.
- $\forall (x, y) \in \mathbb{R}^2$ *with* $f_Y(y) > 0, f_{X|Y}(x|y) = f_X(x)$.
- *For all functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$, if $Z_X = g(X)$ and $Z_Y = h(Y)$, then $Z_X \perp\!\!\!\perp Z_Y$.*
- *For all events A and B where A is defined only in terms of X (i.e., $A = \{X \in D\}$ for some $D \subseteq \mathbb{R}$) and B is defined only in terms of Y (i.e., $B = \{Y \in E\}$ for some $E \subseteq \mathbb{R}$), A and B are independent.*

We omit the proof of this theorem. Notice that the third statement is analogous to Theorem 1.1.7; if X and Y are independent, knowing the outcome for Y gives us *no information* about the probability of any outcome for X.

1.4 Summarizing Distributions

In this section, we discuss some important "summary" features of random variables. These features will be central to our discussion of estimation in the following chapters.

1.4.1 Expected Values

The *expected value* (also known as the *expectation* or *mean*) of a random variable can be thought of as the value we would obtain if we took the average over many, many realizations of that random variable. It is the most commonly used measure of the "center" of a probability distribution.

Definition 1.4.1. *Expected Value*

For a discrete random variable X with PMF f, if $\sum_x |x| f(x) < \infty$,[25] then the expected value of X is:

$$\mathrm{E}[X] = \sum_x x f(x).$$

[25] This regularity condition (and the corresponding one in the continuous case) is known as *absolute convergence*. It is virtually always satisfied in practice, so we omit this technical condition from all subsequent discussion of expectations.

For a continuous random variable X with PDF f, if $\int_{-\infty}^{\infty}|x|f(x)dx < \infty$, then the expected value of X is:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

Functions of a random variable are themselves random variables, so we can calculate their expected values as well. This can be done quite straightforwardly.

Theorem 1.4.1. *Expectation of a Function of a Random Variable*

For all functions $g : \mathbb{R} \to \mathbb{R}$,

- *If X is a discrete random variable with PMF f and $Z = g(X)$, then the expected value of Z is*

$$\mathrm{E}[Z] = \sum_x g(x)f(x).$$

- *If X is a continuous random variable with PDF f and $Z = g(X)$, then the expected value of Z is*

$$\mathrm{E}[Z] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

We omit the proof of this theorem. Note that, when dealing with functions of a random variable, we will often simply write, e.g., $g(X)$ to denote the random variable Z whose value is given by $Z = g(X)$, or $X^2 + 3X$ to denote the random variable W whose value is given by $W = X^2 + 3X$. So we might therefore write $\mathrm{E}[g(X)]$ or $\mathrm{E}[X^2 + 3X]$. In general, any function or expression containing a random variable denotes the random variable whose value is determined by that function or expression.[26]

The following theorem states two basic properties of the expectation operator.

Theorem 1.4.2. *Properties of Expected Values*

For a random variable X,

- $\forall c \in \mathbb{R}, \mathrm{E}[c] = c$.
- $\forall a \in \mathbb{R}, \mathrm{E}[aX] = a\mathrm{E}[X]$.

Proof: A constant c can be considered as a discrete random variable X with the PMF

$$f(x) = \begin{cases} 1 & : x = c \\ 0 & : \text{otherwise} \end{cases}$$

(This is known as a *degenerate distribution*.) Thus,

$$\mathrm{E}[c] = \sum_x xf(x) = cf(c) = c \cdot 1 = c.$$

[26]To be even more formal: given a random variable X whose value is given by $X = \mathcal{X}(\omega)$, where $\mathcal{X} : \Omega \to \mathbb{R}$, and a function $g : \mathcal{X}(\Omega) \to \mathbb{R}$, $g(X)$ denotes the random variable Z whose value is given by $Z = \mathcal{Z}(\omega)$, where $\mathcal{Z} = g \circ \mathcal{X}$.

Now, let X be a random variable and $a \in \mathbb{R}$. Let $g(x) = ax$ and $Z = g(X)$. If X is discrete with PMF f, then by Theorem 1.4.1,

$$\mathrm{E}[aX] = \mathrm{E}[Z] = \sum_x g(x)f(x) = \sum_x axf(x) = a\sum_x xf(x) = a\mathrm{E}[X].$$

Likewise, if X is continuous with PDF f, then by Theorem 1.4.1,

$$\mathrm{E}[aX] = \mathrm{E}[Z] = \int_{-\infty}^{\infty} g(x)f(x)dx = \int_{-\infty}^{\infty} axf(x)dx = a\int_{-\infty}^{\infty} xf(x)dx = a\mathrm{E}[X]. \ \square$$

Let's now consider some examples.

Example 1.4.1. *A Fair Die Roll*

Consider, again, a roll of one fair (six-sided) die. Let X be the value of the outcome of the die roll. Then the expected value of X is

$$\mathrm{E}[X] = \sum_{x=1}^{6} xf(x) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = \frac{7}{2}$$

Note that a random variable does not necessarily take on its expected value with positive probability. In this example, $\Pr(X = \mathrm{E}[X]) = \Pr(X = 7/2) = f(7/2) = f(3.5) = 0$.

Example 1.4.2. *Bernoulli Distribution*

Let X be a Bernoulli random variable. (Recall that we can think of such a random variable as a potentially biased coin flip.) Then

$$\mathrm{E}[X] = \sum_{x=0}^{1} xf(x) = 0(1-p) + 1(p) = p.$$

Notice that this implies a convenient feature of Bernoulli random variables: $\mathrm{E}[X] = \Pr(X = 1)$.

Example 1.4.3. *The Standard Normal Distribution*

Let $X \sim N(0,1)$. Then the expected value of X is

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\left(-e^{-\frac{x^2}{2}}\right)\Bigg|_{-\infty}^{\infty} = 0.$$

We can generalize the concept of expected value to the bivariate case in a couple of ways.[27] The expected value of a random vector (X, Y) is defined simply as the vector of expected values of its components.

Definition 1.4.2. *Expectation of a Bivariate Random Vector*

For a random vector (X, Y), the expected value of (X, Y) is

$$\mathrm{E}[(X,Y)] = (\mathrm{E}[X], \mathrm{E}[Y]).$$

[27] Again, further generalization to the case of three or more random variables can be done analogously.

Furthermore, we can compute the expected value of a function of two random variables, since a function of random variables is itself a random variable.

Theorem 1.4.3. *Expectation of a Function of Two Random Variables*

For any function $h : \mathbb{R}^2 \to \mathbb{R}$*,*

- *If* X *and* Y *are discrete random variables with joint PMF* f *and* $Z = h(X, Y)$*, then the expected value of* Z *is*
$$\mathrm{E}[Z] = \sum_x \sum_y h(x, y) f(x, y).$$
- *If* X *and* Y *are jointly continuous random variables with joint PDF* f *and* $Z = h(X, Y)$*, then the expected value of* Z *is*
$$\mathrm{E}[Z] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y) dy dx.$$

We omit the proof of this theorem. As in the univariate case, note that we will often simply write, e.g., $h(X, Y)$ to denote the random variable Z whose value is given by $Z = h(X, Y)$, or $X^2 + 3Y + XY$ to denote the random variable W whose value is given by $W = X^2 + 3Y + XY$. So we might therefore write $\mathrm{E}[h(X, Y)]$ or $\mathrm{E}[X^2 + 3Y + XY]$. Again, in general, any function or expression containing one or more random variables denotes the random variable whose value is determined by that function or expression.

A consequence of Theorem 1.4.3 is the following generalization of Theorem 1.4.2.

Theorem 1.4.4. *Linearity of Expectations*

Let X *and* Y *be random variables. Then,* $\forall a, b, c \in \mathbb{R}$*,*

$$\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c.$$

Proof: Let X and Y be either discrete random variables with joint PMF f or jointly continuous random variables with joint PDF f,[28] and let $a, b, c \in \mathbb{R}$. Let $h(x, y) = ax + by + c$ and $Z = h(X, Y)$. If X and Y

[28] This theorem also holds when one random variable is discrete and the other continuous, but the proof of that case requires measure theory.

are discrete, then by Theorem 1.4.3,

$$\begin{aligned}
\mathrm{E}\left[aX+bY+c\right] &= \mathrm{E}\left[Z\right] \\
&= \sum_x \sum_y h(x,y) f(x,y) \\
&= \sum_x \sum_y (ax+by+c) f(x,y) \\
&= a\sum_x \sum_y x f(x,y) + b\sum_x \sum_y y f(x,y) + c\sum_x \sum_y f(x,y) \\
&= a\sum_x x \sum_y f(x,y) + b\sum_y y \sum_x f(x,y) + c\sum_x \sum_y f(x,y) \\
&= a\sum_x x f_X(x) + b\sum_y y f_Y(y) + c\sum_x f_X(x) \\
&= a\mathrm{E}\left[X\right] + b\mathrm{E}\left[Y\right] + c\cdot 1 \\
&= a\mathrm{E}\left[X\right] + b\mathrm{E}\left[Y\right] + c.
\end{aligned}$$

Likewise, if X and Y are jointly continuous, then by Theorem 1.4.3,

$$\begin{aligned}
\mathrm{E}\left[aX+bY+c\right] &= \mathrm{E}\left[Z\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f(x,y) dy dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax+by+c) f(x,y) dy dx \\
&= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y) dy dx + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y) dy dx + c\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) dy dx \\
&= a\int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x,y) dy\right] dx + b\int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x,y) dx\right] dy + c\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) dy dx \\
&= a\int_{-\infty}^{\infty} x f_X(x) dx + b\int_{-\infty}^{\infty} y f_Y(y) dy + c\int_{-\infty}^{\infty} f_X(x) dx \\
&= a\mathrm{E}\left[X\right] + b\mathrm{E}\left[Y\right] + c\cdot 1 \\
&= a\mathrm{E}\left[X\right] + b\mathrm{E}\left[Y\right] + c. \ \square
\end{aligned}$$

The expected value is sometimes referred to as the first *moment* of a distribution. We can extend this concept as follows.

Definition 1.4.3. *The j^{th} Moment*

For a random variable X and $j \in \mathbb{N}$, the j^{th} moment of X is:

$$\mu'_j = \mathrm{E}\left[X^j\right].$$

Moments provide summary information about a distribution, describing elements of its shape and location. For $j > 1$, the *j^{th} central moment* generally provides more useful information than the regular j^{th} moment.

Definition 1.4.4. *The j^{th} Central Moment*

For a random variable X and $j \in \mathbb{N}$, the j^{th} central moment of X is:

$$\mu_j = \mathrm{E}\left[(X - \mathrm{E}[X])^j\right].$$

This is referred to as a central moment because it is centered around $\mathrm{E}[X]$. Note that $\mathrm{E}[X]$ is the first moment, *not* the first central moment. The first central moment of any distribution is $\mathrm{E}[X - \mathrm{E}[X]] = \mathrm{E}[X] - \mathrm{E}[X] = 0$.

1.4.2 Variance and Standard Deviation

The 2^{nd} central moment is known as the *variance*. Whereas the expected value of a distribution characterizes its location or center, variance characterizes its variability or spread. Higher variance implies greater unpredictability.[29]

Definition 1.4.5. *Variance*

The variance of a random variable X is

$$V(X) = E\left[(X - \mathrm{E}[X])^2\right].$$

In words, the variance is the average squared deviation from the expected value. The following theorem gives an alternative formula for the variance that is often easier to compute in practice.

Theorem 1.4.5. *Alternative Formula for Variance*

For a random variable X,

$$V(X) = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2.$$

Proof: Let X be a random variable.

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

[29] There are also an infinite number of higher moments, from which one can compute additional features of the shape of a distribution: skewness, kurtosis, etc. In practice, you will never care about anything higher than the 2^{nd} central moment of a distribution (unless you need conditions for asymptotics). If you do, contact the authors and we will issue you a prompt, personal apology.

Notice that $\mathrm{E}[X]$ is a *constant* and is therefore treated as such each time we apply linearity of expectations above. This is why we refer to $\mathrm{E}[\,\cdot\,]$ as an *operator*: unlike when we write, say, $g(X)$, the expectation operator is *not* function of the value of X, and thus $\mathrm{E}[X]$ is *not* a random variable. Rather, it denotes a constant that describes a feature of the distribution of X. The same holds for the variance operator, $V(\,\cdot\,)$. The following theorem states some basic properties of the variance operator. Note carefully how these differ from the properties of expected values (Theorem 1.4.2).

Theorem 1.4.6. *Properties of Variance*

For a random variable X,

- $\forall c \in \mathbb{R}, V(X+c) = V(X)$.
- $\forall a \in \mathbb{R}, V(aX) = a^2 V(X)$.

Proof: Let X be a random variable and let $a, c \in \mathbb{R}$.

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

And

$$\begin{aligned} V(aX) &= \mathrm{E}\left[(aX-\mathrm{E}[aX])^2\right] \\ &= \mathrm{E}\left[(aX-a\mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[a^2(X-\mathrm{E}[X])^2\right] \\ &= a^2\mathrm{E}\left[(X-\mathrm{E}[X])^2\right] \\ &= a^2 V(X). \ \square \end{aligned}$$

The variance is one of the most common measures of the "spread" of a distribution. Another is the *standard deviation*, which is simply the square root of the variance.

Definition 1.4.6. *Standard Deviation*

The standard deviation of a random variable X is

$$\sigma(X) = \sqrt{V(X)}.$$

The following theorem states some basic properties of standard deviation.

Theorem 1.4.7. *Properties of Standard Deviation*

For a random variable X,

- $\forall c \in \mathbb{R}, \sigma(X + c) = \sigma(X)$.
- $\forall a \in \mathbb{R}, \sigma(aX) = |a|\sigma(X)$.

Proof: Simply take the square root of both sides of each equation from Theorem 1.4.6. □

The standard deviation is often preferable to the variance, since it is on same scale as the random variable of interest. We illustrate this with an example.

Example 1.4.4. *A Fair Die Roll*

Consider, again, a roll of one fair (six-sided) die. Let X be the value of the outcome of the die roll. Then

$$V(X) = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2 = \sum_{x=1}^{6}\left(x^2 \cdot \frac{1}{6}\right) - \left(\sum_{x=1}^{6} x \cdot \frac{1}{6}\right)^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{35}{12} \approx 2.92.$$

So

$$\sigma(X) = \sqrt{V(X)} = \frac{\sqrt{105}}{6} \approx 1.71.$$

Now let $Z = 100X$, equivalent to rolling a fair six-sided die with faces labelled 100, 200, 300, 400, 500, and 600. Then

$$V(Z) = V(100X) = 100^2 V(X) = 10000V(X) = \frac{87500}{3} \approx 29200.$$

And

$$\sigma(Z) = \sigma(100X) = |100|\sigma(X) = \frac{50\sqrt{105}}{3} \approx 171.$$

When we scale up the random variable, the standard deviation remains on the same order of magnitude as the range/spread of outcomes. In contrast, the variance of a random variable can "blow up" when we rescale the random variable. Variance is thus more difficult to interpret than standard deviation, since its magnitude does not clearly correspond to the magnitude of the spread of the distribution.

We can learn a lot about a distribution just by knowing its mean and standard deviation. For example, we can put an upper bound on the probability that a draw from the distribution will be more than a given number of standard deviations from the mean.

Theorem 1.4.8. *Chebyshev's Inequality*

Let X be a random variable with finite[30] $\mathrm{E}\left[X\right]$ *and finite $\sigma(X) > 0$. Then $\forall k > 0$,*

$$\Pr\left(|X - \mathrm{E}\left[X\right]| \geq k\sigma(X)\right) \leq \frac{1}{k^2}.$$

[30]Moments can be infinite or undefined. To see this, try taking the expected value of $1/X$, where $X \sim U(0, 1)$. Note also that, if the j^{th} moment of a random variable X is non-finite, then all higher moments (i.e., all k^{th} moments where $k > j$) are also non-finite.

We omit the proof here, but see Goldberger (1991, p. 31) for a simple proof via Markov's Inequality. This theorem will also be important later for showing that estimators converge to the "right" value.

Notice that what we learn about distribution from Chebyshev's Inequality is driven by $\sigma(X)$ rather than $V(X)$. When we know more about the distribution, knowledge of its expected value and standard deviation can be even more informative. For example, in the case of the *normal distribution*, knowing just these two quantities tells us *everything*.

Definition 1.4.7. *The Normal Distribution*

A continuous random variable X follows a normal distribution if it has PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \forall x \in \mathbb{R},$$

for some constants $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$. We write $X \sim N(\mu, \sigma^2)$ to denote that X follows a normal distribution with parameters μ and σ.

The following theorem implies that just knowing the mean and standard deviation of a normal distribution tells us everything about the distribution.

Theorem 1.4.9. *Mean and Standard Deviation of the Normal Distribution*

If $X \sim N(\mu, \sigma^2)$, then

- $\mathrm{E}[X] = \mu$.
- $\sigma(X) = \sigma$.

We omit the proof of this theorem. The parameters μ and σ of a normal distribution are its mean and standard deviation, respectively. A normal distribution is thus uniquely specified by its mean and standard deviation.[31] Furthermore, this is why $N(0, 1)$ is the *standard* normal distribution: it has the "nice" properties of being centered at zero ($\mu = 0$) and having a standard deviation (and variance) of one ($\sigma = \sigma^2 = 1$).

1.4.3 Mean Squared Error

We often want to describe how far off a random variable X is from a certain value c on average. The most commonly used metric is *mean squared error* (MSE), which is the expected value of the squared difference between the observed value of X and c.[32] As we proceed, we will see that MSE has a number of appealing properties for the applied researcher.

[31] Caution: when you see something like $X \sim N(2, 9)$, this means X is normally distributed with mean $\mu = 2$ and *variance* $\sigma^2 = 9$, so $\sigma = \sqrt{9} = 3$. This potentially confusing notation is, unfortunately, the established convention in probability theory.

[32] MSE is most commonly used as a term to quantify the precision of an estimator; as we will discuss in Chapter 2, this is in fact the same definition in a different context. There are alternative metrics, e.g., *mean absolute error* (MAE), which is the expected value of the absolute difference between the observed value of X and c.

Definition 1.4.8. *Mean Squared Error (MSE) about c*

For a random variable X and $c \in \mathbb{R}$, the mean squared error of X about c is $\mathrm{E}\left[(X-c)^2\right]$.

A closely related quantity is the *root mean squared error (RMSE)* about c, $\sqrt{\mathrm{E}\left[(X-c)^2\right]}$. Much as the standard deviation rescales the variance in a way that ensures that it remains on the same scale as the range of outcomes for X, RMSE rescales the MSE so that it remains on the same scale as $X-c$. For these reasons, researchers often prefer to report the RMSE over the MSE.

We can apply Theorem 1.4.5 to derive an alternative formula for MSE.

Theorem 1.4.10. *Alternative Formula for MSE*

For a random variable X and $c \in \mathbb{R}$,

$$\mathrm{E}\left[(X-c)^2\right] = V(X) + (\mathrm{E}[X]-c)^2.$$

Proof: Let X be a random variable and $c \in \mathbb{R}$. Then

$$\begin{aligned}
\mathrm{E}\left[(X-c)^2\right] &= \mathrm{E}\left[X^2 - 2cX + c^2\right] \\
&= \mathrm{E}\left[X^2\right] - 2c\mathrm{E}[X] + c^2 \\
&= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 + \mathrm{E}[X]^2 - 2c\mathrm{E}[X] + c^2 \\
&= \left(\mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2\right) + \left(\mathrm{E}[X]^2 - 2c\mathrm{E}[X] + c^2\right) \\
&= V(X) + (\mathrm{E}[X]-c)^2. \ \square
\end{aligned}$$

Theorem 1.4.10 directly implies that the expected value, $\mathrm{E}[X]$, has an interpretation as the best predictor of X in terms of MSE.

Theorem 1.4.11. *Expected Value Minimizes MSE*

For a random variable X, the value of c that minimizes the mean squared error of X about c is $c = \mathrm{E}[X]$.

Proof: Let X be a random variable. Then

$$\arg\min_{c\in\mathbb{R}} \mathrm{E}\left[(X-c)^2\right] = \arg\min_{c\in\mathbb{R}} \left[V(X) + (\mathrm{E}[X]-c)^2\right] = \arg\min_{c\in\mathbb{R}} \left[(\mathrm{E}[X]-c)^2\right] = \mathrm{E}[X]. \ \square$$

In other words, if we had to pick one number as a prediction of the value of X, the "best" choice (in terms of minimizing MSE) would be $\mathrm{E}[X]$.

Example 1.4.5. *A Fair Coin Flip*

Consider, again, a fair coin flip. Let $X = 0$ if the coin comes up heads and $X = 1$ if the coin comes up tails. What is the *minimum MSE (MMSE)* guess for the value of X? The PMF of X is:

$$f(x) = \begin{cases} \frac{1}{2} & : x \in \{0, 1\} \\ 0 & : \text{otherwise} \end{cases}$$

so the MSE about c is:

$$\mathrm{E}\left[(X-c)^2\right] = \frac{1}{2}(0-c)^2 + \frac{1}{2}(1-c)^2 = \frac{1}{2}\left[c^2 + 1 + c^2 - 2c\right] = \frac{1}{2}\left[1 + 2c^2 - 2c\right] = \frac{1}{2} + c^2 - c.$$

The first-order condition is thus:

$$0 = \frac{d}{dc}\mathrm{E}\left[(X-c)^2\right] = \frac{d}{dc}\left(\frac{1}{2} + c^2 - c\right) = 2c - 1,$$

which is solved by $c = \frac{1}{2}$. Note that this is just $\mathrm{E}[X]$ (see Example 1.4.2).

Figure 1.3.1 illustrates this solution.

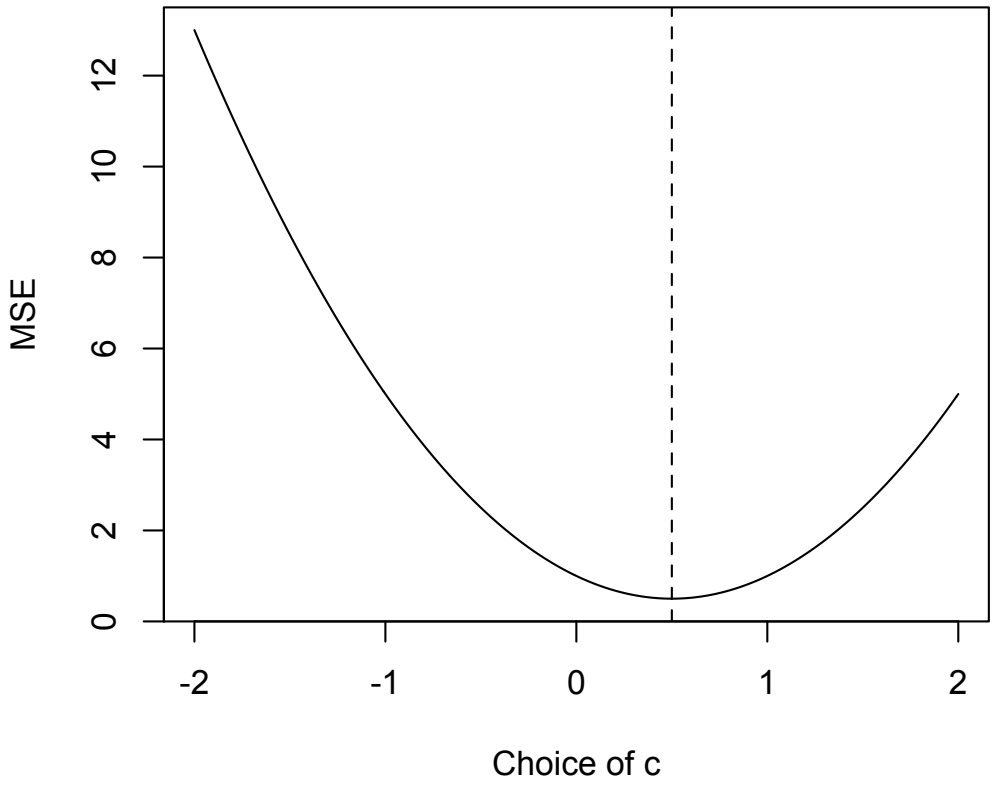


Figure 1.3.1 *MMSE Solution for a Fair Coin Flip*

1.4.4 Covariance and Correlation

The generalization of variance to the bivariate case is the *covariance*. Covariance measures the extent to which two random variables "move together." If X and Y have positive covariance, that means that, when the value of X is larger, the value of Y tends to be larger. If X and Y have negative covariance, then the opposite is true: when the value of X is larger, the value of Y tends to be smaller.

Definition 1.4.9. *Covariance*

The covariance of two random variables X and Y is

$$\mathrm{Cov}\,(X,Y) = \mathrm{E}\left[\,(X - \mathrm{E}\,[X])\,(Y - \mathrm{E}\,[Y])\,\right].$$

As with expected value and variance, note that $\mathrm{Cov}\,(\,\cdot\,,\,\cdot\,)$ is what we call an operator, not a function of the values its arguments take on, so $\mathrm{Cov}\,(X,Y)$ is a constant, not a random variable. As with variance, there is an alternative formula for covariance that is generally easier to compute in practice.

Theorem 1.4.12. *Alternative Formula for Covariance*

For random variables X and Y,

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

Proof: Let X and Y be random variables.

$$\begin{aligned}
\mathrm{Cov}\,(X,Y) &= \mathrm{E}\left[\,(X - \mathrm{E}\,[X])\,(Y - \mathrm{E}\,[Y])\,\right] \\
&= \mathrm{E}\left[XY - X\mathrm{E}\,[Y] - Y\mathrm{E}\,[X] + \mathrm{E}\,[X]\mathrm{E}\,[Y]\right] \\
&= \mathrm{E}\,[XY] - \mathrm{E}\,[X]\mathrm{E}\,[Y] - \mathrm{E}\,[X]\mathrm{E}\,[Y] + \mathrm{E}\,[X]\mathrm{E}\,[Y] \\
&= \mathrm{E}\,[XY] - \mathrm{E}\,[X]\mathrm{E}\,[Y]. \ \square
\end{aligned}$$

We now derive some important properties of variance and covariance. The following theorem generalizes Theorem 1.4.6.

Theorem 1.4.13. *Variance Rule*

Let X and Y be random variables. Then

$$V(X+Y) = V(X) + 2\mathrm{Cov}\,(X,Y) + V(Y).$$

More generally, $\forall a, b, c \in \mathbb{R}$,

$$V(aX + bY + c) = a^2V(X) + 2ab\mathrm{Cov}\,(X,Y) + b^2V(Y).$$

Note that we do *not* have linearity of variances. To remember this theorem, it may be helpful to recall the algebraic formula $(x+y)^2 = x^2 + 2xy + y^2$. As the following proof shows, the similarity is not merely coincidental.

Proof: Let X and Y be random variables. Then

$$\begin{aligned}
V(X+Y) &= \mathrm{E}\left[(X+Y)^2\right] - \mathrm{E}\left[X+Y\right]^2 \\
&= \mathrm{E}\left[X^2 + 2XY + Y^2\right] - \Big(\mathrm{E}\left[X\right] + \mathrm{E}\left[Y\right]\Big)^2 \\
&= \mathrm{E}\left[X^2\right] + 2\mathrm{E}\left[XY\right] + \mathrm{E}\left[Y^2\right] - \mathrm{E}\left[X\right]^2 - 2\mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] - \mathrm{E}\left[Y\right]^2 \\
&= \Big(\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2\Big) + 2\Big(\mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]\Big) + \Big(\mathrm{E}\left[Y^2\right] - \mathrm{E}\left[Y\right]^2\Big) \\
&= V(X) + 2\mathrm{Cov}\left(X,Y\right) + V(Y).
\end{aligned}$$

The proof of the more general version is left as an exercise to the reader. □

Theorem 1.4.14. *Properties of Covariance*
For random variables X, Y, and Z,

- $\forall c, d \in \mathbb{R}, \mathrm{Cov}\left(c, X\right) = \mathrm{Cov}\left(X, c\right) = \mathrm{Cov}\left(c, d\right) = 0.$
- $\mathrm{Cov}\left(X, Y\right) = \mathrm{Cov}\left(Y, X\right).$
- $\mathrm{Cov}\left(X, X\right) = V(X).$
- $\forall a, b, c, d \in \mathbb{R}, \mathrm{Cov}\left(aX + c, bY + d\right) = ab\mathrm{Cov}\left(X, Y\right).$
- $\mathrm{Cov}\left(X, Y + Z\right) = \mathrm{Cov}\left(X, Y\right) + \mathrm{Cov}\left(X, Z\right).$

Proof: Let X, Y, and Z be random variables and let $a, b, c, d \in \mathbb{R}$. Then

$$\mathrm{Cov}\left(c, X\right) = \mathrm{E}\left[cX\right] - \mathrm{E}\left[c\right]\mathrm{E}\left[X\right] = c\mathrm{E}\left[X\right] - c\mathrm{E}\left[X\right] = 0.$$

$$\mathrm{Cov}\left(X, c\right) = \mathrm{E}\left[Xc\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[c\right] = c\mathrm{E}\left[X\right] - c\mathrm{E}\left[X\right] = 0.$$

$$\mathrm{Cov}\left(c, d\right) = \mathrm{E}\left[cd\right] - \mathrm{E}\left[c\right]\mathrm{E}\left[d\right] = cd - cd = 0.$$

$$\mathrm{Cov}\left(X, Y\right) = \mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] = \mathrm{E}\left[YX\right] - \mathrm{E}\left[Y\right]\mathrm{E}\left[X\right] = \mathrm{Cov}\left(Y, X\right).$$

$$\mathrm{Cov}\left(X, X\right) = \mathrm{E}\left[XX\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[X\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2 = V(X).$$

$$\begin{aligned}
\text{Cov}\,(aX+c, bY+d) &= \text{E}\left[(aX+c-\text{E}\,[aX+c])(bY+d-\text{E}\,[bY+d])\right] \\
&= \text{E}\left[(aX+c-a\text{E}\,[X]-c)(bY+d-b\text{E}\,[Y]-d)\right] \\
&= \text{E}\left[a(X-\text{E}\,[X])b(Y-\text{E}\,[Y])\right] \\
&= ab\text{E}\left[(X-\text{E}\,[X])(Y-\text{E}\,[Y])\right] = ab\text{Cov}\,(X,Y).
\end{aligned}$$

$$\begin{aligned}
\text{Cov}\,(X, Y+Z) &= \text{E}\left[X(Y+Z)\right] - \text{E}\,[X]\text{E}\,[Y+Z] \\
&= \text{E}\,[XY+XZ] - \text{E}\,[X]\Big(\text{E}\,[Y]+\text{E}\,[Z]\Big) \\
&= \text{E}\,[XY] + \text{E}\,[XZ] - \text{E}\,[X]\text{E}\,[Y] - \text{E}\,[X]\text{E}\,[Z] \\
&= \Big(\text{E}\,[XY] - \text{E}\,[X]\text{E}\,[Y]\Big) + \Big(\text{E}\,[XZ] - \text{E}\,[X]\text{E}\,[Z]\Big) \\
&= \text{Cov}\,(X,Y) + \text{Cov}\,(X,Z). \ \square
\end{aligned}$$

Notice that variance is effectively a special case of covariance: the covariance of a random variable with itself is its variance. Thus, covariance is indeed a generalization of variance. Another measure of the relationship between two random variables is their *correlation*.

Definition 1.4.10. *Correlation*

The correlation of two random variables X and Y with $\sigma(X) > 0$ and $\sigma(Y) > 0$ is

$$\rho(X,Y) = \frac{\text{Cov}\,(X,Y)}{\sigma(X)\sigma(Y)}.$$

Much like standard deviation rescales variance, correlation rescales covariance to make its interpretation clearer. Correlation measures linear dependence and is bounded in $[-1, 1]$, where a correlation of 1 represents perfect positive linear dependence, a correlation of -1 represents perfect negative linear dependence, and a correlation of 0 implies no linear relationship. We formalize these facts (except for the last one) in the following theorem.

Theorem 1.4.15. *Correlation and Linear Dependence*

For random variable X and Y,

- $\rho(X,Y) \in [-1,1]$.
- $\rho(X,Y) = 1 \iff \exists\, a, b \in \mathbb{R}$ *with* $b > 0$ *such that* $Y = a + bX$.
- $\rho(X,Y) = -1 \iff \exists\, a, b \in \mathbb{R}$ *with* $b > 0$ *such that* $Y = a - bX$.

We omit the proof of this theorem.[33] The following theorem states some other important properties of correlation.

Theorem 1.4.16. *Properties of Correlation*

For random variables X, Y, and Z,

- $\rho(X,Y) = \rho(Y,X)$.
- $\rho(X,X) = 1$.
- $\rho(aX+c, bY+d) = \rho(X,Y)$, $\forall a,b,c,d \in \mathbb{R}$ *such that either* $a,b>0$ *or* $a,b<0$.
- $\rho(aX+c, bY+d) = -\rho(X,Y)$, $\forall a,b,c,d \in \mathbb{R}$ *such that either* $a<0<b$ *or* $b<0<a$.

Proof: Let X and Y be random variables and let $a,b,c,d \in \mathbb{R}$ with $a,b \neq 0$. Then

$$\rho(X,Y) = \frac{\text{Cov}\,(X,Y)}{\sigma(X)\sigma(Y)} = \frac{\text{Cov}\,(Y,X)}{\sigma(Y)\sigma(X)} = \rho(Y,X).$$

$$\rho(X,X) = \frac{\text{Cov}\,(X,X)}{\sigma(X)\sigma(X)} = \frac{V(X)}{\sigma(X)^2} = \frac{V(X)}{\left(\sqrt{V(X)}\right)^2} = \frac{V(X)}{V(X)} = 1.$$

$$\rho(aX+c, bY+d) = \frac{\text{Cov}\,(aX+c, bY+d)}{\sigma(aX+c)\sigma(bY+d)} = \frac{ab\text{Cov}\,(X,Y)}{|a||b|\sigma(X)\sigma(Y)} = \frac{ab}{|ab|}\rho(X,Y).$$

If $a,b>0$ or $a,b<0$ (i.e., a and b have the same sign), then $\frac{ab}{|ab|} = 1$, so

$$\rho(aX+c, bY+d) = \rho(X,Y).$$

And if $a<0<b$ or $b<0<a$ (i.e., a and b have opposite signs), then $\frac{ab}{|ab|} = -1$, so

$$\rho(aX+c, bY+d) = -\rho(X,Y). \ \square$$

[33] The fact that correlation is bounded in $[-1,1]$ is equivalent to the well-known *Cauchy-Schwarz Inequality*, which, in one form, states that, for random variables X and Y, $\text{Cov}\,(X,Y)^2 \leq V(X)V(Y)$. This equivalence is shown as follows:

$$\text{Cov}\,(X,Y)^2 \leq V(X)V(Y) \iff \frac{\text{Cov}\,(X,Y)^2}{V(X)V(Y)} \leq 1 \iff \sqrt{\frac{\text{Cov}\,(X,Y)^2}{V(X)V(Y)}} \leq 1 \iff \frac{\sqrt{\text{Cov}\,(X,Y)^2}}{\sqrt{V(X)}\sqrt{V(Y)}} \leq 1$$
$$\iff \frac{|\text{Cov}\,(X,Y)|}{\sigma(X)\sigma(Y)} \leq 1 \iff -1 \leq \frac{\text{Cov}\,(X,Y)}{\sigma(X)\sigma(Y)} \leq 1 \iff -1 \leq \rho(X,Y) \leq 1.$$

1.4.5 Independence

We can now derive some additional properties of independent random variables related to the features of distributions discussed above.

Theorem 1.4.17. *Implications of Independence (Part II)*

If X and Y are independent random variables, then

- $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.
- *Covariance is zero:* $\mathrm{Cov}(X, Y) = 0$.
- *Correlation is zero:* $\rho(X, Y) = 0$.
- *Variances are additive:* $V(X + Y) = V(X) + V(Y)$.

Proof: Let X and Y be either two discrete independent random variables with joint PMF $f(x, y)$ or two jointly continuous independent random variables with joint PDF $f(x, y)$. Then, $\forall x, y \in \mathbb{R}$, $f(x, y) = f_X(x)f_Y(y)$. So if X and Y are discrete, then

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_x \sum_y xyf(x, y) \\
&= \sum_x \sum_y xyf_X(x)f_Y(y) \\
&= \sum_x xf_X(x) \sum_y yf_Y(y) \\
&= \mathrm{E}[X]\mathrm{E}[Y].
\end{aligned}$$

Likewise, if X and Y are jointly continuous, then

$$\begin{aligned}
\mathrm{E}[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dydx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dydx \\
&= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy \\
&= \mathrm{E}[X]\mathrm{E}[Y].
\end{aligned}$$

Thus,

$$\mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = \mathrm{E}[X]\mathrm{E}[Y] - \mathrm{E}[X]\mathrm{E}[Y] = 0,$$

and

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{0}{\sqrt{V(X)}\sqrt{V(Y)}} = 0.$$

Finally, by Theorem 1.4.13,

$$V(X,Y) = V(X) + 2\text{Cov}\,(X,Y) + V(Y) = V(X) + 2(0) + V(Y) = V(X) + V(Y). \ \square$$

Recall that random variables X and Y are independent if and only if knowing the outcome for one provides no information about the probability of any outcome for the other. In other words, independence means there is *no relationship* between the outcome for X and the outcome for Y. Thus, it is not surprising that no relationship implies no linear relationship. However, unlike in Part I (Theorem 1.3.4), the converses of the statements in Theorem 1.4.17 do *not* necessarily hold, as illustrated in the following example.

Example 1.4.6. *Zero Covariance Does Not Imply Independence*

Let X be a discrete random variable with marginal PMF:

$$f_X(x) = \begin{cases} \frac{1}{3} & : x \in \{-1, 0, 1\} \\ 0 & : \text{otherwise} \end{cases}$$

and let $Y = X^2$, so Y has marginal PMF:

$$f_Y(y) = \begin{cases} \frac{1}{3} & : y = 0 \\ \frac{2}{3} & : y = 1 \\ 0 & : \text{otherwise} \end{cases}$$

and the joint PMF of X and Y is:

$$f(x,y) = \begin{cases} \frac{1}{3} & : (x,y) \in \{(0,0), (1,1), (-1,1)\} \\ 0 & : \text{otherwise} \end{cases}$$

Clearly, X and Y are not independent, e.g., $f(0,1) = 0 \neq \frac{2}{9} = \frac{1}{3} \cdot \frac{2}{3} = f_X(0) f_Y(1)$. Yet

$$\begin{aligned} \text{Cov}\,(X,Y) &= \text{E}\,[XY] - \text{E}\,[X]E[Y] \\ &= \sum_x \sum_y xy f(x,y) - \sum_x x f_X(x) \sum_y y f_Y(y) \\ &= \left[-1 \cdot 1 \cdot \frac{1}{3} + 0 \cdot 0 \cdot \frac{1}{3} + 1 \cdot 1 \cdot \frac{1}{3} \right] - \left[\left(-1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} \right) \left(0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} \right) \right] \\ &= 0 - 0 \cdot \frac{2}{3} \\ &= 0. \end{aligned}$$

So we have $\text{Cov}\,(X,Y) = 0$, but $X \not\perp\!\!\!\perp Y$.

1.4.6 Conditional Expectations

Just as we were able to compute the conditional probability of an event and the conditional PMF or PDF of a random variable, we can compute the *conditional expectation* of a random variable, i.e., the expected value of a random variable given that some other random variable takes on a certain value. To calculate a conditional expectation of a random variable, we simply replace the (marginal) PMF/PDF with the appropriate conditional PMF/PDF in the formula for expected value.

Definition 1.4.11. *Conditional Expectation*

For two discrete random variables X and Y with joint PMF f, the conditional expectation of Y given $X = x$ is:

$$\mathrm{E}\left[Y|X = x\right] = \sum_y y f_{Y|X}(y|x), \forall x \in \mathbb{R} \text{ with } f_X(x) > 0.$$

For two jointly continuous random variables X and Y with joint PDF f, the conditional expectation of Y given $X = x$ is:

$$\mathrm{E}\left[Y|X = x\right] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy, \forall x \in \mathbb{R} \text{ with } f_X(x) > 0.$$

As with regular expectations (see Theorem 1.4.1 and Theorem 1.4.3), we can apply conditional expectations to functions of one or more random variables.

Theorem 1.4.18. *Conditional Expectation of a Function of Random Variables*

For any function $h : \mathbb{R}^2 \to \mathbb{R}$,

- *If X and Y are discrete random variables with joint PMF f and Z = h(X,Y), then the conditional expectation of Z given $X = x$ is:*

$$\mathrm{E}\left[Z|X = x\right] = \sum_y h(x, y) f_{Y|X}(y|x), \forall x \in \mathbb{R} \text{ with } f_X(x) > 0.$$

- *If X and Y are jointly continuous random variables with joint PDF f and $Z = h(X, Y)$, then the conditional expectation of Z given $X = x$ is:*

$$\mathrm{E}\left[Z|X = x\right] = \int_{-\infty}^{\infty} h(x, y) f_{Y|X}(y|x)dy, \forall x \in \mathbb{R} \text{ with } f_X(x) > 0.$$

We omit the proof of this theorem.

Everything else carries through just as though we were operating on a univariate PMF/PDF. For example, we can define the *conditional variance* of Z given $X = x$.

Definition 1.4.12. *Conditional Variance*

For any function $h : \mathbb{R}^2 \to \mathbb{R}$, if X and Y are random variables and $Z = h(X, Y)$, then the conditional variance of Z given $X = x$ is:

$$V(Z|X = x) = \mathrm{E}\left[\left(h(X, Y) - \mathrm{E}\left[h(X, Y)|X = x\right]\right)^2 |X = x\right], \forall x \in \mathbb{R} \text{ with } f_X(x) > 0.$$

As with regular variance, we can derive an alternative formula for conditional variance that is generally easier to work with in practice.

Theorem 1.4.19. *Alternative Formula for Conditional Variance*

For any function $h : \mathbb{R}^2 \to \mathbb{R}$, if X and Y are random variables and $Z = h(X, Y)$, then

$$V(Z|X = x) = \mathrm{E}\left[h(X,Y)^2|X = x\right] - \mathrm{E}\left[h(X,Y)|X = x\right]^2, \forall x \in \mathbb{R} \text{ with } f_X(x) > 0.$$

The proof is essentially the same as the proof of Theorem 1.4.5.

We now turn to a central topic of this book: the *conditional expectation function (CEF)*. The CEF is the function that takes as an input x and returns the conditional expectation of Y given $X = x$.

Definition 1.4.13. *Conditional Expectation Function (CEF)*

For random variables X and Y with joint PMF/PDF f, the conditional expectation function of Y given $X = x$ is:

$$G_Y(x) = \mathrm{E}\left[Y|X = x\right], \forall x \in \mathbb{R} \text{ with } f_X(x) > 0.$$

A few remarks on notation are in order here. We will generally simply write $\mathrm{E}\left[Y|X = x\right]$ to denote the CEF rather than $G_Y(x)$. The above definition is merely meant to emphasize that, when we use the term CEF, we are referring to the *function* that yields $\mathrm{E}\left[Y|X = x\right]$ rather than the value of $\mathrm{E}\left[Y|X = x\right]$ at some specific x. It is also intended to clarify that the CEF, $\mathrm{E}\left[Y|X = x\right]$, is a univariate function of x. It is *not* a function of the random variable Y.

$G_Y(X)$ is a function of the random variable X and is therefore itself a random variable whose value depends on the value of X. I.e., when X takes on the value x, the random variable $Z = G_Y(X)$ takes on the value $G_Y(x) = \mathrm{E}\left[Y|X = x\right]$. We write $\mathrm{E}\left[Y|X\right]$ to denote $G_Y(X)$, since $\mathrm{E}\left[Y|X = X\right]$ would be confusing. Also note that we can analogously define the *conditional variance function*, $H_Y(x) = V(Y|X = x)$, which we will generally write simply as $V(Y|X = x)$, and write $V(Y|X)$ to denote the random variable $W = H_Y(X)$.

Just like unconditional expectations, conditional expectations are linear.

Theorem 1.4.20. *Linearity of Conditional Expectations*

For random variables X and Y and functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$,

$$\mathrm{E}\left[g(X)Y + h(X)|X\right] = g(X)\mathrm{E}\left[Y|X\right] + h(X).$$[34]

Proof: This is an immediate consequence of Theorem 1.4.18. □

[34] Note that this is equivalent to the statement: $\forall x \in \mathbb{R}$ with $f_X(x) > 0$,

$$\mathrm{E}\left[g(X)Y + h(X)|X = x\right] = g(x)\mathrm{E}\left[Y|X = x\right] + h(x).$$

(This follows from Definition 1.3.1.) From this point on, we will generally state theorems involving conditional expectations using the $\mathrm{E}\left[\cdot|X\right]$ notation for the sake of simplicity.

We now apply this concept to our familiar example of flipping a coin and rolling a four- or six-sided die.

Example 1.4.7. *Flipping a Coin and Rolling a Die*

Consider, again, the generative process from Example 1.1.3. Recall from Example 1.3.1 that the conditional PMF of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{4} & : x = 0, y \in \{1,2,3,4\} \\ \frac{1}{6} & : x = 1, y \in \{1,2,3,4,5,6\} \\ 0 & : otherwise \end{cases}$$

Thus, the CEF of Y given $X = x$ is:

$$\begin{aligned} \mathrm{E}\left[Y|X=x\right] &= \sum_y y f_{Y|X}(y|x) \\ &= \begin{cases} \sum_{y=1}^{4} y \cdot \frac{1}{4} & : x = 0 \\ \sum_{y=1}^{6} y \cdot \frac{1}{6} & : x = 1 \end{cases} \\ &= \begin{cases} \frac{5}{2} & : x = 0 \\ \frac{7}{2} & : x = 1 \end{cases} \end{aligned}$$

Likewise, the conditional PMF of X given $Y = y$ is:

$$f_{X|Y}(x|y) = \begin{cases} \frac{3}{5} & : x = 0, y \in \{1,2,3,4\} \\ \frac{2}{5} & : x = 1, y \in \{1,2,3,4\} \\ 1 & : x = 1, y \in \{5,6\} \\ 0 & : \text{otherwise} \end{cases}$$

so the CEF of X given $Y = y$ is:

$$\begin{aligned} \mathrm{E}\left[X|Y=y\right] &= \sum_x x f_{X|Y}(x|y) \\ &= \begin{cases} 0 \cdot \frac{3}{5} + 1 \cdot \frac{2}{5} & : y \in \{1,2,3,4\} \\ 0 \cdot 0 + 1 \cdot 1 & : y \in \{5,6\} \end{cases} \\ &= \begin{cases} \frac{2}{5} & : y \in \{1,2,3,4\} \\ 1 & : y \in \{5,6\} \end{cases} \end{aligned}$$

We can now state one of the most important theorems in this book, the *Law of Iterated Expectations*. The Law of Iterated Expectations is important because it allows us to move between conditional expectations and unconditional expectations. Very often, conditional expectations are easier to work with—in such cases, we can treat some random variables as fixed, which allows for more tractable calculations.

Theorem 1.4.21. *Law of Iterated Expectations*

For any function $h : \mathbb{R}^2 \to \mathbb{R}$, if X and Y are random variables and $Z = h(X, Y)$, then

$$\mathrm{E}\,[Z] = \mathrm{E}_X\Big[\mathrm{E}\,[Z|X]\Big],$$

where E_X refers to the expectation over X.[35]

Proof: Let X and Y be either two discrete random variables with joint PMF $f(x, y)$ or two jointly continuous random variables with joint PDF $f(x, y)$. If X and Y are discrete, then

$$\begin{aligned}
\mathrm{E}\,[Z] &= \sum_x \sum_y h(x,y) f(x,y) \\
&= \sum_x \sum_y h(x,y) f_X(x) f_{Y|X}(y|x) \\
&= \sum_x \left[\sum_y h(x,y) f_{Y|X}(y|x) \right] f_X(x) \\
&= \sum_x \mathrm{E}\,[Z|X=x] f_X(x) \\
&= \mathrm{E}_X\Big[\mathrm{E}\,[Z|X]\Big].
\end{aligned}$$

Likewise, if X and Y are jointly continuous, then

$$\begin{aligned}
\mathrm{E}\,[Z] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f(x,y) dy dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f_X(x) f_{Y|X}(y|x) dy dx \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} h(x,y) f_{Y|X}(y|x) dy \right] f_X(x) dx \\
&= \int_{-\infty}^{\infty} \mathrm{E}\,[Z|X=x] f_X(x) dx \\
&= \mathrm{E}_X\Big[\mathrm{E}\,[Z|X]\Big]. \ \square
\end{aligned}$$

Put simply, the Law of Iterated Expectations implies that the unconditional expectation can be represented as a weighted average of conditional expectations, where the weights are proportional to the probability

[35] We will often simply write E instead of E_X. Both are correct; E_X is just clearer.

distribution of the variable being conditioned on. The Law of Iterated Expectations will be invoked frequently in proofs throughout the remainder of this book. The Law of Iterated Expectations also directly yields the *Law of Total Variance*.

Theorem 1.4.22. *Law of Total Variance*

For random variables X and Y,

$$V(Y) = \mathrm{E}_X\Big[V(Y|X)\Big] + V_X\Big(\mathrm{E}\left[Y|X\right]\Big).$$

Proof: Let X and Y be random variables. Then

$$\begin{aligned}
V(Y) &= \mathrm{E}\left[Y^2\right] - \mathrm{E}\left[Y\right]^2 \\
&= \mathrm{E}_X\left[\mathrm{E}\left[Y^2|X\right]\right] - \mathrm{E}_X\Big[\mathrm{E}\left[Y|X\right]\Big]^2 \\
&= \mathrm{E}_X\left[V(Y|X) + \mathrm{E}\left[Y|X\right]^2\right] - \mathrm{E}_X\Big[\mathrm{E}\left[Y|X\right]\Big]^2 \\
&= \mathrm{E}_X\Big[V(Y|X)\Big] + \left(\mathrm{E}_X\Big[\mathrm{E}\left[Y|X\right]^2\Big] - \mathrm{E}_X\Big[\mathrm{E}\left[Y|X\right]\Big]^2\right) \\
&= \mathrm{E}_X\Big[V(Y|X)\Big] + V_X\Big(E[Y|X]\Big). \ \square
\end{aligned}$$

Practically speaking, this theorem allows us to decompose the variability of a random variable Y into the average variability "within" values of X and the variability "across" values of X.

Returning to the CEF, the following properties of deviations from the CEF will be necessary to demonstrate its significance.

Theorem 1.4.23. *Properties of Deviations from the CEF*

Let X and Y be random variables and let $\epsilon = Y - \mathrm{E}\left[Y|X\right]$. Then

- $\mathrm{E}\left[\epsilon\right] = 0$.
- $\mathrm{E}\left[\epsilon|X\right] = 0$.
- *For any function $g : \mathbb{R} \to \mathbb{R}$,* $\mathrm{Cov}\Big(g(X), \epsilon\Big) = 0$.
- $V[\epsilon] = \mathrm{E}\Big[V(Y|X)\Big]$.

Proof: Let X and Y be random variables and let $\epsilon = Y - \mathrm{E}\left[Y|X\right]$. Applying the Law of Iterated Expectations,

$$\mathrm{E}\left[\epsilon\right] = \mathrm{E}\Big[Y - \mathrm{E}\left[Y|X\right]\Big] = \mathrm{E}\left[Y\right] - \mathrm{E}\Big[\mathrm{E}\left[Y|X\right]\Big] = \mathrm{E}\left[Y\right] - \mathrm{E}\left[Y\right] = 0.$$

Noting that $\mathrm{E}[Y|X]$ is a function solely of X and applying Theorem 1.4.20,

$$\mathrm{E}[\epsilon|X] = \mathrm{E}\Big[Y - \mathrm{E}[Y|X]|X\Big] = \mathrm{E}[Y|X] - \mathrm{E}[Y|X] = 0.$$

Let $g : \mathbb{R} \to \mathbb{R}$. Then

$$\begin{aligned}
\mathrm{Cov}(g(X), \epsilon) &= \mathrm{E}\Big[g(X)\epsilon\Big] - \mathrm{E}\Big[g(X)\Big]\mathrm{E}[\epsilon] \\
&= \mathrm{E}\left[g(X)\Big(Y - \mathrm{E}[Y|X]\Big)\right] - \mathrm{E}\Big[g(X)\Big](0) \\
&= \mathrm{E}\Big[g(X)Y - g(X)\mathrm{E}[Y|X]\Big] \\
&= \mathrm{E}\Big[g(X)Y\Big] - \mathrm{E}\Big[g(X)\mathrm{E}[Y|X]\Big] \\
&= \mathrm{E}\Big[g(X)Y\Big] - \mathrm{E}\Big[\mathrm{E}\Big[g(X)Y|X\Big]\Big] \\
&= \mathrm{E}\Big[g(X)Y\Big] - \mathrm{E}\Big[g(X)Y\Big] \\
&= 0.
\end{aligned}$$

Finally,

$$\begin{aligned}
V(\epsilon) &= \mathrm{E}[\epsilon^2] - \mathrm{E}[\epsilon]^2 \\
&= \mathrm{E}\left[\Big(Y - \mathrm{E}[Y|X]\Big)^2\right] - 0^2 \\
&= \mathrm{E}\Big[Y^2 - 2Y\mathrm{E}[Y|X] + \mathrm{E}[Y|X]^2\Big] \\
&= \mathrm{E}[Y^2] - 2\mathrm{E}\Big[Y\mathrm{E}[Y|X]\Big] + \mathrm{E}\Big[\mathrm{E}[Y|X]^2\Big] \\
&= \mathrm{E}_X\Big[\mathrm{E}[Y^2|X]\Big] - 2\mathrm{E}_X\Big[\mathrm{E}\Big[Y\mathrm{E}[Y|X]|X\Big]\Big] + \mathrm{E}\Big[\mathrm{E}[Y|X]^2\Big] \\
&= \mathrm{E}_X\Big[\mathrm{E}[Y^2|X]\Big] - 2\mathrm{E}_X\Big[\mathrm{E}[Y|X]\mathrm{E}[Y|X]\Big] + \mathrm{E}_X\Big[\mathrm{E}[Y|X]^2\Big] \\
&= \mathrm{E}_X\Big[\mathrm{E}[Y^2|X]\Big] - 2\mathrm{E}_X\Big[\mathrm{E}[Y|X]^2\Big] + \mathrm{E}_X\Big[\mathrm{E}[Y|X]^2\Big] \\
&= \mathrm{E}_X\Big[\mathrm{E}[Y^2|X]\Big] - \mathrm{E}_X\Big[\mathrm{E}[Y|X]^2\Big] \\
&= \mathrm{E}_X\Big[\mathrm{E}[Y^2|X] - \mathrm{E}[Y|X]^2\Big] \\
&= \mathrm{E}\left[V(Y|X)\right]. \ \square
\end{aligned}$$

We can now prove a crucial fact. Suppose we knew the full joint CDF of X and Y, and then someone gave us a randomly drawn value of X. What would be the guess of Y that would have the lowest MSE? Formally, what function $g : \mathbb{R} \to \mathbb{R}$ minimizes $\mathrm{E}\left[(Y - g(X))^2\right]$? The answer is given by the CEF. The function $g(X)$ that best approximates Y is the CEF.

Theorem 1.4.24. *The CEF is the MMSE Predictor*

For random variables X and Y, the CEF, $\mathrm{E}[Y|X]$, is the best (MMSE) predictor of Y given X.

Proof: Choose any function $g(X)$ to approximate Y. Let $U = Y - g(X)$. By the definition of MMSE, our goal is to choose $g(X)$ to minimize $\mathrm{E}[U^2]$. Let $\epsilon = Y - \mathrm{E}[Y|X]$ and $W = \mathrm{E}[Y|X] - g(X)$, so that: $U = \epsilon + W$. (Note that W is a function of X, so we can treat it like a constant in expectation conditioned on X.) Our goal is then to show that, by choosing $g(X) = \mathrm{E}[Y|X]$, we will minimize $\mathrm{E}[U^2]$. Now,

$$\begin{aligned}
\mathrm{E}[U^2|X] &= \mathrm{E}\left[(\epsilon + W)^2|X\right] \\
&= \mathrm{E}\left[\epsilon^2 + 2\epsilon W + W^2|X\right] \\
&= \mathrm{E}\left[\epsilon^2|X\right] + 2W\mathrm{E}[\epsilon|X] + W^2 \\
&= \mathrm{E}\left[\epsilon^2|X\right] + 0 + W^2 \\
&= \mathrm{E}\left[\epsilon^2|X\right] - \mathrm{E}[\epsilon|X]^2 + W^2 \\
&= V[Y|X] + W^2,
\end{aligned}$$

where the third and fifth lines follow from Theorem 1.4.23. Applying Law of Iterated Expectations:

$$\mathrm{E}[U^2] = \mathrm{E}\left[V[Y|X] + W^2\right] = \mathrm{E}\left[V[Y|X]\right] + \mathrm{E}\left[W^2\right].$$

$\mathrm{E}\left[V[Y|X]\right]$ does not depend on the choice of $g(X)$. And $\mathrm{E}\left[W^2\right] \geq 0$, with equality if $g(X) = \mathrm{E}[Y|X]$. Therefore, choosing $g(X) = \mathrm{E}[Y|X]$ minimizes MSE. □

1.4.7 The Best Linear Predictor

What if we were to restrict ourselves to just linear functions? Among functions of the form $g(X) = a + bX$, what function yields the best (MMSE) prediction of Y given X. By choosing the a and b that minimize MSE, we obtain the *best linear predictor* (BLP) of Y given X.

Theorem 1.4.25. *Best Linear Predictor (BLP)*

For random variables X and Y, the best linear predictor of Y given X is $g(X) = \alpha + \beta X$, where

$$\begin{aligned}
\alpha &= \mathrm{E}[Y] - \frac{\mathrm{Cov}(X, Y)}{V[X]}\mathrm{E}[X], \\
\beta &= \frac{\mathrm{Cov}(X, Y)}{V[X]}.
\end{aligned}$$

Proof: We apply the first order condition to minimizing $\mathrm{E}[U^2]$, where $U = Y - (a + bX)$. We want to choose a and b to solve the system

$$\begin{cases} 0 = \dfrac{\partial \mathrm{E}[U^2]}{\partial a} \\ \\ 0 = \dfrac{\partial \mathrm{E}[U^2]}{\partial b} \end{cases}$$

By linearity of expectations and the chain rule, this becomes:

$$\begin{cases} 0 = \dfrac{\partial \mathrm{E}[U^2]}{\partial a} = \mathrm{E}\left[\dfrac{\partial U^2}{\partial a}\right] = \mathrm{E}\left[2U\dfrac{\partial U}{\partial a}\right] = -2\mathrm{E}[U] \\ \\ 0 = \dfrac{\partial \mathrm{E}[U^2]}{\partial b} = \mathrm{E}\left[\dfrac{\partial U^2}{\partial b}\right] = \mathrm{E}\left[2U\dfrac{\partial U}{\partial b}\right] = -2\mathrm{E}[UX] \end{cases}$$

Now we just solve $0 = \mathrm{E}[Y - (a + bX)]$ and $0 = \mathrm{E}[(Y - (a + bX))X] = 0$. From the first equation,

$$0 = \mathrm{E}[Y - (a + bX)] = \mathrm{E}[Y] - a - b\mathrm{E}[X],$$

so $a = \mathrm{E}[Y] - b\mathrm{E}[X]$. Then from the second equation,

$$\begin{aligned} 0 &= \mathrm{E}[(Y - (a + bX))X] \\ &= \mathrm{E}\left[YX - aX - bX^2\right] \\ &= \mathrm{E}[XY] - a\mathrm{E}[X] - b\mathrm{E}\left[X^2\right] \\ &= \mathrm{E}[XY] - (\mathrm{E}[Y] - b\mathrm{E}[X])\,\mathrm{E}[X] - b\mathrm{E}\left[X^2\right] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] + b\mathrm{E}[X]^2 - b\mathrm{E}\left[X^2\right] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] - b\left(\mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2\right) \\ &= \mathrm{Cov}(X, Y) - bV(X). \end{aligned}$$

Solving for b, we obtain:

$$b = \frac{\mathrm{Cov}(X, Y)}{V(X)} = \beta.$$

Finally, substituting back into $a = \mathrm{E}[Y] - b\mathrm{E}[X]$ yields:

$$a = \mathrm{E}[Y] - \frac{\mathrm{Cov}(X, Y)}{V(X)}\mathrm{E}[X] = \alpha. \ \square$$

Note that α is the y-intercept of the BLP, and β is its slope.[36] We note two important corollaries:

- The BLP is also the best linear approximation to the CEF (i.e., setting $a = \alpha$ and $b = \beta$ minimizes $\mathrm{E}\left[\left[\mathrm{E}\left[Y|X\right] - (a + bX)\right]^2\right]$. The proof is obtained by the same means as above.
- If the CEF is linear, the CEF *is* the BLP.

The CEF and the BLP are very important, as each permits a principled and simple summary of the way in which the best prediction of one variable is related to the value of another variable. And both the CEF and BLP are generalizations of the simple expected value.

But why would we use the BLP when the CEF is "better"? One reason is that, whereas the CEF might be infinitely complex, the BLP is characterized just by two numbers, α and β. The BLP is a simple approximation for the CEF, and one that operates on the same principle—find the function that minimizes MSE—but with the further restriction that the function must be linear.

We now consider an example to show how the CEF and BLP allow us to summarize bivariate relationships.

Example 1.4.8. *Plotting the CEF and BLP*

Let X and Y be random variables with $X \sim U(0,1)$ and $Y = 10X^2 + W$, where $W \sim N(0,1)$ and $X \perp\!\!\!\perp W$. We derive the CEF of Y given X as follows.

$$\mathrm{E}\left[Y|X\right] = \mathrm{E}\left[10X^2 + W|X\right]$$

By linearity of expectations:

$$\mathrm{E}\left[10X^2 + W|X\right] = 10\mathrm{E}\left[X^2|X\right] + \mathrm{E}\left[W|X\right] = 10\mathrm{E}\left[X^2|X\right] + 0 = 10X^2.$$

Thus, the CEF of Y given X is $\mathrm{E}\left[Y|X\right] = 10X^2$.

We now derive the BLP of Y given X. The slope of the BLP is:

$$\begin{aligned}
\beta &= \frac{\mathrm{Cov}\left(X,Y\right)}{V[X]} \\
&= \frac{\mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2} \\
&= \frac{\mathrm{E}\left[X\left(10X^2 + W\right)\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[10X^2 + W\right]}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2} \\
&= \frac{\mathrm{E}\left[10X^3 + XW\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[10X^2 + W\right]}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2}.
\end{aligned}$$

[36] Readers with prior training in statistics or econometrics may recognize the expression for the BLP as resembling the OLS regression solution. This is not an accident. We will explain this similarity in Chapter 3.

By linearity of expectations,

$$\begin{aligned}
\beta &= \frac{\mathrm{E}\left[10X^3 + XW\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[10X^2 + W\right]}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2} \\
&= \frac{10\mathrm{E}\left[X^3\right] + \mathrm{E}\left[XW\right] - \mathrm{E}\left[X\right]\left(10\mathrm{E}\left[X^2\right] + \mathrm{E}\left[W\right]\right)}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2} \\
&= \frac{10\mathrm{E}\left[X^3\right] + \mathrm{E}\left[XW\right] - 10\mathrm{E}\left[X\right]\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[W\right]}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2} \\
&= \frac{10\left(\mathrm{E}\left[X^3\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[X^2\right]\right) + \left(\mathrm{E}\left[XW\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[W\right]\right)}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2} \\
&= \frac{10\left(\mathrm{E}\left[X^3\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[X^2\right]\right) + \mathrm{Cov}\left(X, W\right)}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2}.
\end{aligned}$$

By independence,

$$\begin{aligned}
\beta &= \frac{10\left(\mathrm{E}\left[X^3\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[X^2\right]\right) + \mathrm{Cov}\left(X, W\right)}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2} \\
&= \frac{10\left(\mathrm{E}\left[X^3\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[X^2\right]\right) + 0}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2} \\
&= \frac{10\left(\mathrm{E}\left[X^3\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[X^2\right]\right)}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2}.
\end{aligned}$$

$X \sim U(0,1)$, so its PDF is $f_X(x) = 1$, $\forall x \in [0,1]$, and $f(x) = 0$ otherwise, so

$$\begin{aligned}
\beta &= \frac{10\left(\mathrm{E}\left[X^3\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[X^2\right]\right)}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2} \\
&= \frac{10\left[\int_0^1 (x^3 \cdot 1)dx - \left(\int_0^1 (x \cdot 1)dx\right)\left(\int_0^1 (x^2 \cdot 1)dx\right)\right]}{\int_0^1 (x^2 \cdot 1)dx - \left(\int_0^1 (x \cdot 1)dx\right)^2} \\
&= \frac{10\left[\frac{1}{4}x^4|_0^1 - \left(\frac{1}{2}x^2|_0^1\right)\left(\frac{1}{3}x^3|_0^1\right)\right]}{\frac{1}{3}x^3|_0^1 - \left(\frac{1}{2}x^2|_0^1\right)^2}
\end{aligned}$$

$$= \frac{10\left[\frac{1}{4}(1-0) - \frac{1}{2}(1-0)\frac{1}{3}(1-0)\right]}{\frac{1}{3}(1-0) - \left[\frac{1}{2}(1-0)\right]^2}$$

$$= \frac{10\left(\frac{1}{4} - \frac{1}{6}\right)}{\frac{1}{3} - \left(\frac{1}{2}\right)^2} = \frac{10\left(\frac{1}{12}\right)}{\frac{1}{12}}$$

$$= 10.$$

Finally, the intercept is

$$\alpha = \mathrm{E}\left[Y\right] - \beta \mathrm{E}\left[X\right] = \mathrm{E}\left[10X^2 + W\right] - \beta \mathrm{E}\left[X\right].$$

By linearity of expectations,

$$\begin{aligned}
\alpha &= \mathrm{E}\left[10X^2 + W\right] - \beta \mathrm{E}\left[X\right] \\
&= 10\mathrm{E}\left[X^2\right] + \mathrm{E}\left[W\right] - \beta \mathrm{E}\left[X\right] \\
&= 10\mathrm{E}\left[X^2\right] + 0 - \beta \mathrm{E}\left[X\right] \\
&= 10\int_0^1 (x^2 \cdot 1)dx - 10\int_0^1 (x \cdot 1)dx \\
&= 10 \cdot \frac{x^3}{3}\Big|_0^1 - 10 \cdot \frac{x^2}{2}\Big|_0^1 \\
&= \frac{10}{3} - \frac{10}{2} \\
&= -\frac{5}{3}.
\end{aligned}$$

Thus, the BLP of Y given X is $g(X) = -\frac{5}{3} + 10X$.

Figure 1.3.2 plots 1200 random draws of (X, Y) and superimposes the graphs of the CEF (in red) and the BLP (in blue).

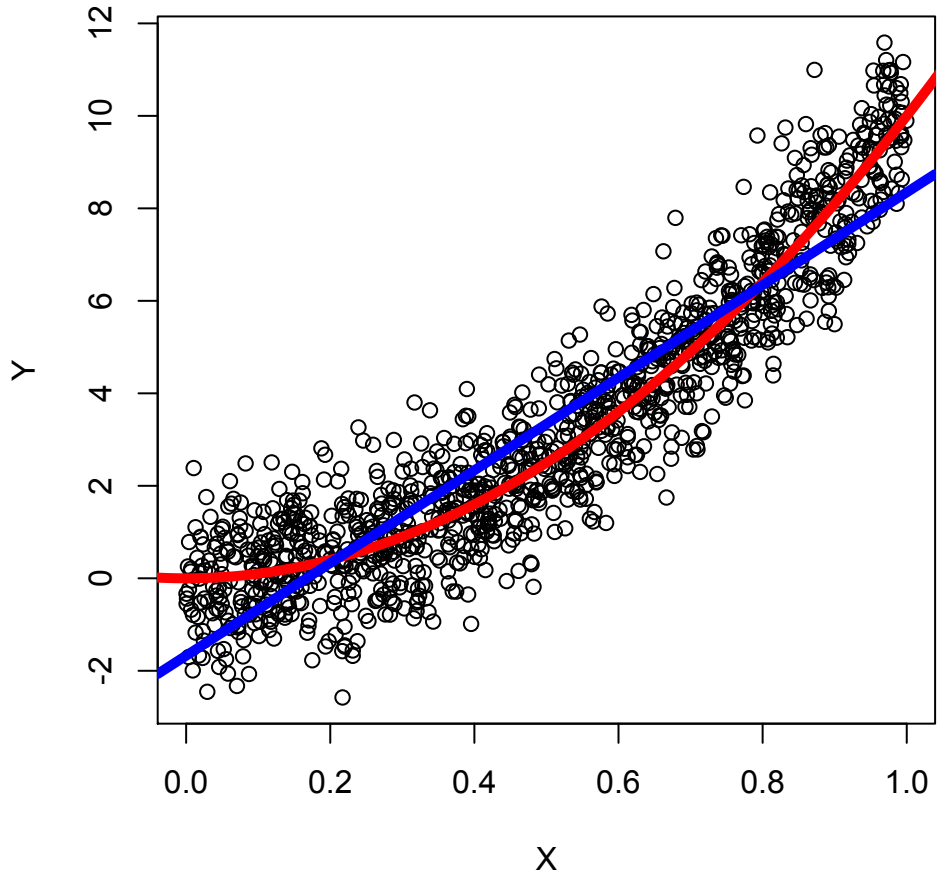


Figure 1.3.2. *Plotting the CEF and BLP*

Note that the BLP approximates the CEF reasonably well over the range of the data. While this is not always the case, it is very often the case in the social and health sciences. The BLP is thus a good "first approximation" in a very literal sense, in that it is an approximation with a first-order polynomial.[37]

1.4.8 The CEF and BLP under Independence

We can now derive some additional properties of independent random variables as they relate to conditional expectations, the CEF, and the BLP.

Theorem 1.4.26. *Implications of Independence (Part III)*

If X and Y are independent random variables, then

[37] We will discuss approximation of the CEF with higher-order polynomials in Section 3.4.2.

- $\mathrm{E}[Y|X] = \mathrm{E}[Y]$.
- $V(Y|X) = V(Y)$.
- *For all functions* $g : \mathbb{R} \to \mathbb{R}$ *and* $h : \mathbb{R} \to \mathbb{R}$, $\mathrm{E}[g(Y)|h(X)] = \mathrm{E}[g(Y)]$.
- *The BLP of* Y *given* X *is* $\mathrm{E}[Y]$.
- *For all functions* $g : \mathbb{R} \to \mathbb{R}$ *and* $h : \mathbb{R} \to \mathbb{R}$, *the BLP of* $g(Y)$ *given* $h(X)$ *is* $\mathrm{E}[g(Y)]$.

We omit a proof for Theorem 1.4.26, noting that it directly follows from Theorems 1.3.4 and 1.4.17.

1.4.9 Multivariate Generalizations

We conclude this chapter by briefly outlining the generalizations of some of the above concepts to the case where we have more than two random variables, with the goal of defining the CEF and BLP when we have more than one explanatory variable. We provide no examples and only one proof in this section, as our goal is simply to illustrate that the concepts we have presented in detail in the bivariate case are easily extended. Essentially all omitted definitions, properties, and theorems carry through, *mutatis mutandis*.

Consider the random vector $(X_1, X_2, ..., X_K, Y)$.[38] As in the bivariate case, the distribution of a random vector is completely described by its joint CDF.

Definition 1.4.14. *Joint CDF (Multivariate Case)*

For random variables X_1, X_2, ..., X_K, *and* Y, *the joint CDF* F *of* X_1, X_2, ...,X_K, *and* Y *is:*

$$F(x_1, x_2, ..., x_K, y) = \Pr(X_1 \leq x_1, X_2 \leq x_2, ..., X_K \leq x_K, Y \leq y), \forall (x_1, x_2, ..., x_K, y) \in \mathbb{R}^{K+1}.$$

As before, the joint CDF at $(x_1, x_2, ..., x_K, y)$ gives the probability of observing $X_1 \leq x_1$, $X_2 \leq x_2$, ..., $X_K \leq x_K$, and $Y \leq y$ in a single realization of $(X_1, X_2, ..., X_K, Y)$.

In the case where X_1, X_2, ..., X_K, and Y are all discrete random variables, we can define the joint PMF of X_1, X_2, ..., X_K, and Y just as before.

Definition 1.4.15. *Joint PMF (Multivariate Case)*

For discrete random variables X_1, X_2, ..., X_K, *and* Y, *the joint PMF* f *of* X_1, X_2, ...,X_K, *and* Y *is:*

$$f(x_1, x_2, ..., x_K, y) = \Pr(X_1 = x_1, X_2 = x_2, ..., X_K = x_K, Y = y), \forall (x_1, x_2, ..., x_K, y) \in \mathbb{R}^{K+1}.$$

Again, as before, the joint PMF at $(x_1, x_2, ..., x_K, y)$ is simply the probability that $X_1 = x_1$, $X_2 = x_2$, ..., $X_K = x_K$, and $Y = y$ in a single realization of $(X_1, X_2, ..., X_K, Y)$.

In the case where X_1, X_2, ..., X_K and Y are all continuous random variables, we must first define joint continuity before we can define the joint PDF.

[38] Although we do not need a variable specially designated as Y at this point, it will be useful to maintain this notation once we consider generalizations to the CEF and BLP.

Definition 1.4.16. *Jointly Continuous Random Variables (Multivariate Case)*

The random variables X_1, X_2, ..., X_K, and Y are jointly continuous if there exists a nonnegative function $f : \mathbb{R}^{K+1} \to \mathbb{R}$ such that the joint CDF F of X_1, X_2, ..., X_K, and Y is:

$$F(x_1, x_2, ..., x_K, y) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_K} \int_{-\infty}^{y} f(u_1, u_2, ..., u_K, v) dv du_K ... du_2 du_1,$$

$\forall (x_1, x_2, ..., x_K, y) \in \mathbb{R}^{K+1}$. *The function f is the joint PDF.*

As in the bivariate case, we define the joint PDF more explicitly as follows.

Definition 1.4.17. *Joint PDF (Multivariate Case)*

For continuous random variables X_1, X_2, ..., X_K, and Y with joint CDF F, the joint PDF f of X_1, X_2, ..., X_K, and Y is:

$$f(x_1, x_2, ..., x_K, y) = \left. \frac{\partial^K F(u_1, u_2, ..., u_K)}{\partial u_1 \partial u_2 \cdots \partial u_K \partial v} \right|_{u_1 = x_1, u_2 = x_2, ..., u_K = x_K, v = y}, \forall (x_1, x_2, ..., x_K, y) \in \mathbb{R}^{K+1}.$$

The intuition is the same as in the bivariate case, and event probabilities are again obtained by integrating the PDF over the region that corresponds to that event.

We can also define marginal and conditional PMFs/PDFs in the multivariate case.

Definition 1.4.18. *Marginal PMF/PDF (Multivariate Case)*

For discrete random variables X_1, X_2, ..., X_K, and Y with joint PMF f, the marginal PMF of Y is:

$$f_Y(y) = \Pr(Y = y) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_K} f(x_1, x_2, ..., x_K, y), \forall y \in \mathbb{R}.$$

For jointly continuous random variables X_1, X_2, ..., X_K, and Y with joint PDF f, the marginal PDF of Y is:

$$f_Y(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, ..., x_K, y) dx_1 dx_2 ... dx_K, \forall y \in \mathbb{R}.$$

As before, the idea behind the marginal PMF/PDF is simple: we just sum (or integrate) over all possible outcomes for the X variables to obtain the univariate PMF/PDF of Y, that is, the PMF/PDF of Y ignoring the existence of the Xs. Note that this is actually just one of many types of marginal distributions we can obtain in the multivariate case. We could similarly sum (or integrate) over just Y to obtain the marginal (joint) distribution of the Xs, ignoring the existence of Y. Indeed, we could choose any arbitrary $k \leq K$ of these random variables to "marginalize out," thereby obtaining the marginal (joint) distribution of the remaining variables.

At this point, we must introduce a little vector notation for the sake of clarity. Let $\mathbf{X} = (X_1, X_2, ..., X_K)$, and denote specific outcome values for $\mathbf{X}$ by $\mathbf{x} = (x_1, x_2, ..., x_K)$.[39] Let $f_{\mathbf{X}}(\mathbf{x})$ denote the marginal (joint) distribution of $\mathbf{X}$, i.e., in the discrete case, $f_{\mathbf{X}}(\mathbf{x}) = \sum_y f(\mathbf{x}, y)$, and in the continuous case, $f_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x}, y) dy$.

[39] All vectors in this book shall be denoted by boldface letters.

Definition 1.4.19. *Conditional PMF/PDF (Multivariate Case)*

For discrete random variables X_1, X_2, ..., X_K, and Y with joint PMF f, the conditional PMF of Y given $\mathbf{X} = \mathbf{x}$ is:

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \Pr(Y = y|\mathbf{X} = \mathbf{x}) = \frac{\Pr(Y = y, \mathbf{X} = \mathbf{x})}{\Pr(\mathbf{X} = \mathbf{x})} = \frac{f(\mathbf{x}, y)}{f_{\mathbf{X}}(\mathbf{x})}, \forall (\mathbf{x}, y) \in \mathbb{R}^{K+1} \text{ with } f_{\mathbf{X}}(\mathbf{x}) > 0.$$

For jointly continuous random variables X_1, X_2, ..., X_K, and Y with joint PDF f, the conditional PDF of Y given $\mathbf{X} = \mathbf{x}$ is:

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \frac{f(\mathbf{x}, y)}{f_{\mathbf{X}}(\mathbf{x})}, \forall (\mathbf{x}, y) \in \mathbb{R}^{K+1} \text{ with } f_{\mathbf{X}}(\mathbf{x}) > 0.$$

We can now define conditional expectations and the CEF in the multivariate case.

Definition 1.4.20. *Conditional Expectations (Multivariate Case)*

For discrete random variables X_1, X_2, ..., X_K, and Y with joint PMF f, the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$ is:

$$\mathrm{E}[Y|\mathbf{X} = \mathbf{x}] = \sum_y y f_{Y|\mathbf{X}}(y|\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^K \text{ with } f_{\mathbf{X}}(\mathbf{x}) > 0.$$

For jointly continuous random variables X_1, X_2, ..., X_K, and Y with joint PDF f, the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$ is

$$\mathrm{E}[Y|\mathbf{X} = \mathbf{x}] = \int_{-\infty}^{\infty} y f_{Y|\mathbf{X}}(y|\mathbf{x}) dy, \forall \mathbf{x} \in \mathbb{R}^K \text{ with } f_{\mathbf{X}}(\mathbf{x}) > 0.$$

Definition 1.4.21. *CEF (Multivariate Case)*

For random variables X_1, X_2, ..., X_K, and Y with joint PMF/PDF f, the conditional expectation function of Y given $\mathbf{X} = \mathbf{x}$ is:

$$G_Y(\mathbf{x}) = \mathrm{E}[Y|\mathbf{X} = \mathbf{x}], \forall \mathbf{x} \in \mathbb{R}^K \text{ with } f_{\mathbf{X}}(\mathbf{x}) > 0.$$

As before, this definition is merely meant to emphasize that the CEF refers to the *function* of $\mathbf{x}$ that yields $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}]$; in general, we will simply write $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}]$ to denote the CEF. The same notational conventions noted in the bivariate case apply: $\mathrm{E}[Y|\mathbf{X}]$ denotes $G_Y(\mathbf{X})$; the conditional variance function is $H_Y(\mathbf{x}) = V(Y|\mathbf{X} = \mathbf{x})$, generally written simply as $V(Y|\mathbf{X} = \mathbf{x})$; and $V(Y|\mathbf{X})$ denotes $H_Y(\mathbf{X})$.

In the multivariate case, the CEF is the *still* the best (MMSE) predictor of Y given X_1, X_2, ..., X_K. That is, suppose we knew the full joint CDF of X_1, X_2, ..., X_K, and Y, and then someone gave us a randomly drawn value of $(X_1, X_2, ..., X_K)$. What would be the guess of Y that would have the lowest MSE? Formally, what function $g : \mathbb{R}^K \to \mathbb{R}$ minimizes $\mathrm{E}\left[(Y - g(X_1, X_2, ..., X_K))^2\right]$? Again, the answer is given by the CEF.

Theorem 1.4.27. *The CEF is the MMSE Predictor*

For random variables X_1, X_2, ..., X_K, and Y, the CEF, $\mathrm{E}[Y|\mathbf{X}]$, is the best (MMSE) predictor of Y given $\mathbf{X}$.

The proof is the same as in the bivariate case.

Finally, we can describe the BLP of Y given X_1, X_2, ..., X_K. As before, the BLP is the *linear* function that minimizes MSE.

Definition 1.4.22. *BLP (Multivariate Case)*

For random variables X_1, X_2, ..., X_K, and Y, the best linear predictor of Y given $\mathbf{X}$ (i.e., the MMSE predictor of Y given $\mathbf{X}$ among functions of the form $g(\mathbf{X}) = b_0 + b_1X_1 + b_2X_2... + b_KX_K)$) is $g(\mathbf{X}) = \beta_0 + \beta_1X_1 + \beta_2X_2... + \beta_KX_K$, where

$$(\beta_0, \beta_1, \beta_2, ..., \beta_K) = \underset{(b_0,b_1,b_2,...,b_K)\in\mathbb{R}^{K+1}}{\arg\min} \mathrm{E}\left[[Y - (b_0 + b_1X_1 + b_2X_2... + b_KX_K)]^2\right].$$

That is, among functions of the form $g(\mathbf{X}) = b_0 + b_1X_1 + ... + b_KX_K$, what function yields the best (MMSE) prediction of Y given X_1, X_2, ..., X_K? Formally, what values of b_0, b_1, b_2, ..., b_K minimize $\mathrm{E}\left[(Y - (b_0 + b_1X_1 + b_2X_2 + ... + b_KX_K))^2\right]$? In the multivariate case, we no longer have the nice, simple formulas $\alpha = \mathrm{E}[Y] - \frac{\mathrm{Cov}(X,Y)}{V[X]}\mathrm{E}[X]$ and $\beta = \frac{\mathrm{Cov}(X,Y)}{V[X]}$ for the solution. Instead, the full solution involves either matrix algebra or multiple steps. We shall return to this in Chapter 3.

The BLP has many of the same properties in the multivariate case that it had in the bivariate case. As before:

- The BLP is also the best linear approximation to the CEF (i.e., setting $(b_0, b_1, b_2, ..., b_K) = (\beta_0, \beta_1, \beta_2, ..., \beta_K)$ minimizes $\mathrm{E}\left[[\mathrm{E}[Y|\mathbf{X}] - (b_0 + b_1X_1 + b_2X_2 + ... + b_KX_K)]^2\right]$. The proof is obtained by the same means as above.
- If the CEF is linear, the CEF *is* the BLP.

One remarkable feature of the BLP is that it facilitates easy interpretation. First, β_0, is the intercept (sometimes known as the constant), and represents the value the BLP would take on if all of the variables were held at zero—i.e., $g(0, 0, ..., 0) = \beta_0$. To see this, we need only note that $g(0, 0, ..., 0) = \beta_0 + \beta_10 + \beta_20 + ...\beta_K0 = \beta_0$.

The remaining coefficients of the BLP represent partial derivatives. E.g., β_1 represents how much the BLP of Y would change if we moved one unit on X_1, holding all else equal. This is easy to see, by considering a set of arbitrary values $(x_1, x_2, ..., x_K)$, so that $g(x_1, x_2, ..., x_K) = \beta_0 + \beta_1x_1 + \beta_2x_2... + \beta_Kx_k$. We need only note that $g(x_1 + 1, x_2, ..., x_K) = \beta_0 + \beta_1(x_1 + 1) + \beta_2x_2... + \beta_Kx_k$. Therefore, $g(x_1 + 1, x_2, ..., x_K) - g(x_1, x_2, ..., x_K) = \beta_1$.

Put another way, β_1 is the slope of the BLP with respect to X_1, *conditional* on the values of all of the other variables. If we held X_2, X_3, ..., X_K fixed at some numbers, then how would the BLP change by moving X_1 by one? The same intuition holds for β_2, β_3, and so on. We formalize this in the following theorem.

Theorem 1.4.28. *Coefficients of the BLP are Partial Derivatives*

For random variables X_1, X_2, ..., X_K, and Y, if $g(\mathbf{X})$ is the best linear predictor of Y given $\mathbf{X}$, then $\forall k \in \{1, 2, ..., K\}$,

$$\beta_k = \frac{\partial g(\mathbf{X})}{\partial X_k}.$$

Proof: This follows immediately from linearity:

$$\frac{\partial g(X_1, X_2, ..., X_K)}{\partial X_k} = \frac{\partial \left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_K X_K\right)}{\partial X_k} = \beta_k. \ \square$$

When the CEF is well approximated by the BLP, then these properties of the BLP afford us a simple way of describing features of the CEF.

2 Learning from Random Samples

Ignorance gives one a large range of probabilities.

— GEORGE ELIOT

In Chapter 1, we discussed the properties of random variables, a theoretical construct. In this chapter and the next one, we turn to the question of what we can learn about a random variable by observing repeated draws of that random variable. Under suitable regularity conditions, with infinite draws of a random variable, we can learn *everything* about its CDF. With a finite n (number of draws of the random variable), we can produce *estimates* of everything about its CDF, and, furthermore, estimate the uncertainty of those estimates. The former is discussed in Section 2.1, while the latter is the subject of Section 2.2.

2.1 Estimation

If we observed just one draw of a random variable X, we probably couldn't learn too much about its distribution. Unless X had no variance, any guess about a feature of its distribution (e.g., CDF, expected value) would be subject to a great deal of chance variability. We would not be able to *estimate* a feature of the distribution of the X with much reliability.

But what if we observed multiple draws of a random variable? As our sample, or collection of observations from the same random variable X, grows larger and larger, we are able to estimate the features of X with more and more precision. We will first focus on the properties of one *sample statistic*: the sample mean. The sample mean of draws from X may be viewed as an estimator of $\mathrm{E}[X]$, and has properties that we will now characterize.

2.1.1 The Sample Mean

Suppose we have n *independent and identically distributed (i.i.d.)* observations of a random variable X.

Definition 2.1.1. *Independent and Identically Distributed (i.i.d.)*

Let X_1, X_2, ..., X_n be random variables with CDFs F_1, F_2, ..., F_n, respectively. Let F_A denote the joint CDF of the random variables with indices in the set A. Then X_1, X_2, ..., X_n are independently and identically distributed if they satisfy the following:

- *Mutually independent:* $\forall A \subseteq \{1, 2, ..., n\}$, $\forall (x_1, x_2, ..., x_n) \in \mathbb{R}^n$, $F_A((x_i)_{i \in A}) = \prod_{i \in A} F_i(x_i)$
- *Identically distributed:* $\forall i, j \in \{1, 2, ..., n\}$ *and* $\forall x \in \mathbb{R}$, $F_i(x) = F_j(x)$.

Note that mutual independence implies that all observations are pairwise independent: $\forall i \neq j$, $X_i \perp\!\!\!\perp X_j$, but the converse does not generally hold.

In other words, we take a draw of the random variable X. Then we take another draw of X, in such a manner that our second draw does not depend on the outcome of the first. We repeat this process until we have n draws. So, we have n identical random variables that generate n values. We can subscript each of these n random variables, as shown above, as well as aggregate them to form the random vector $\mathbf{X} = (X_1, ..., X_n)^T$. Under the assumption of i.i.d. sampling from X, the collection of values that we see are each produced by an identical random process. The actual values that we observe are denoted by $\mathbf{x} = (x_1, ..., x_n)^T$.[40]

Note that when we have i.i.d. random variables X_1, X_2, ..., X_n, we use the unsubscripted letter X to denote the random variable whose distribution the i.i.d. draws all share. So, for example, $\mathrm{E}[X]$ denotes the expected value common to X_1, X_2, ..., X_n, i.e., $\mathrm{E}[X] = \mathrm{E}[X_1] = \mathrm{E}[X_2] = ... = \mathrm{E}[X_n]$. For reasons that will become clear in Section 2.1.3, we will often refer to $\mathrm{E}[X]$ as the *population mean* of X.

A *sample statistic* summarizes the values of X_1, X_2, ..., X_n. For now, we shall focus on one crucial sample statistic, the *sample mean*.

Definition 2.1.2. *Sample Mean*

For i.i.d. random variables X_1, X_2, ..., X_n, the sample mean is:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n}(X_1 + X_2 + ... + X_n).$$

Importantly, the sample mean is a function of the random variables X_1, X_2, ..., X_n, so it is *itself* a random variable! Depending on the outcomes of the random variables X_1, X_2, ..., X_n, $\overline{X}$ will take on different values.

We now state some important properties of $\overline{X}$.

Theorem 2.1.1. *The Sample Mean is Unbiased for the Population Mean*[41]

For i.i.d. random variables X_1, X_2, ..., X_n,

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}[X].$$

Proof: Let X_1, X_2, ..., X_n be i.i.d. random variables. Then

$$\begin{aligned}\mathrm{E}[\overline{X}] &= \mathrm{E}\left[\frac{1}{n}(X_1 + X_2 + ... + X_n)\right] \\ &= \frac{1}{n}\mathrm{E}\left[X_1 + X_2 + ... + X_n\right] \\ &= \frac{1}{n}(\mathrm{E}[X_1] + \mathrm{E}[X_2] + ... + \mathrm{E}[X_n])\end{aligned}$$

[40] We are expressing the random vector induced by random sampling as a column vector to clarify that each row represents a separate observation. Later, when we address i.i.d. draws from a random vector, this notational distinction will be helpful.

[41] Note that the phrase "is unbiased for" is equivalent to stating "is an unbiased estimator of." We formally define unbiasedness in Section 2.1.5

$$= \frac{1}{n}\left(\mathrm{E}\left[X\right] + \mathrm{E}\left[X\right] + ... + \mathrm{E}\left[X\right]\right)$$

$$= \frac{1}{n} n \mathrm{E}\left[X\right] = \mathrm{E}\left[X\right]. \ \square$$

The variance of $\overline{X}$ is known as the *sampling variance* of the sample mean.

Theorem 2.1.2. *Sampling Variance of the Sample Mean*

For i.i.d. random variables X_1, X_2, ..., X_n, the sampling variance of $\overline{X}$ is:

$$V(\overline{X}) = \frac{V(X)}{n}.$$

Proof: Let X_1, X_2, ..., X_n be i.i.d. random variables. Then

$$V(\overline{X}) = V\left(\frac{1}{n}\left(X_1 + X_2 + ... + X_n\right)\right)$$

$$= \frac{1}{n^2} V\left(X_1 + X_2 + ... + X_n\right)$$

$$= \frac{1}{n^2}\left(V(X_1) + V(X_2) + ... + V(X_n)\right)$$

$$= \frac{1}{n^2}\left(V(X) + V(X) + ... + V(X)\right)$$

$$= \frac{1}{n^2} n V(X)$$

$$= \frac{V(X)}{n}. \ \square$$

Notice that the sampling variance $V(\overline{X})$ decreases as n increases. This fact, along with Theorem 2.1.1, underlies an important result that we will present shortly. First, however, we must establish the following theorem.

Theorem 2.1.3. *Chebyshev's Inequality for the Sample Mean*

Let X_1, X_2, ..., X_n be i.i.d. random variables with finite $\mathrm{E}\left[X\right]$ and finite $V(X) > 0$. Then, $\forall k > 0$,

$$\Pr\left(|\overline{X} - \mathrm{E}\left[X\right]| \geq k\right) \leq \frac{V(X)}{k^2 n}.$$

Proof: This follows directly from the previous two theorems and Chebyshev's Inequality (Theorem 1.4.8). Let X_1, X_2, ..., X_n be i.i.d. random variables with finite $\mathrm{E}\left[X\right]$ and finite $V(X) > 0$. Let $k > 0$ and let $k' = {}^{k}/_{\sigma(\overline{X})}$, so that $k = k'\sigma(\overline{X})$. Then

$$\Pr\left(|\overline{X} - \mathrm{E}\left[X\right]| \geq k\right) = \Pr\left(|\overline{X} - \mathrm{E}\left[\overline{X}\right]| \geq k'\sigma(\overline{X})\right) \leq \frac{1}{k'^2} = \frac{\sigma(\overline{X})^2}{k^2} = \frac{V(\overline{X})}{k^2} = \frac{V(X)}{k^2 n}. \ \square$$

Theorem 2.1.3 allows us to prove a crucial theorem in statistics, the *Weak Law of Large Numbers (WLLN)*. To state the WLLN, we will also require the following definition.

Definition 2.1.3. *Convergence in Probability*

Let $(X_1, X_2, X_3, ...)$ be a sequence of random variables and let $c \in \mathbb{R}$. Then X_n converges in probability to c if, $\forall k > 0$,

$$\lim_{n\to\infty} \Pr\left(|X_n - c| \geq k\right) = 0,$$

or equivalently,[42]

$$\lim_{n\to\infty} \Pr\left(|X_n - c| < k\right) = 1.$$

We write $X_n \xrightarrow{p} c$ to denote that X_n converges in probability to c.[43]

Intuitively, saying that X_n converges in probability to c means that as n gets large, it becomes increasingly likely that X_n will be "close" to c. More specifically, for any $p \in [0, 1]$ and any $k > 0$, there is always an N large enough that, for every $n \geq N$, the probability that X_n will be within k of c is at least p. As $n \to \infty$, it becomes extremely likely that X_n will be extremely close to c.

We can now state the Weak Law of Large Numbers.

Theorem 2.1.4. *The Weak Law of Large Numbers (WLLN)*

Let $(X_1, X_2, X_3, ...)$ be a sequence of i.i.d. random variables[44] *with finite $\mathrm{E}[X]$ and finite $V(X) > 0$, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then*

$$\overline{X}_n \xrightarrow{p} \mathrm{E}[X].$$

Proof: Let $(X_1, X_2, X_3, ...)$ be a sequence of i.i.d. random variables with finite $\mathrm{E}[X]$ and finite $V(X) > 0$, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. By non-negativity of probability and Theorem 2.1.3, $\forall k > 0$,

$$0 \leq \Pr\left(\left|\overline{X}_n - \mathrm{E}[X]\right| \geq k\right) \leq \frac{V(X)}{k^2 n}.$$

And

$$\lim_{n\to\infty} 0 = \lim_{n\to\infty} \frac{V(X)}{k^2 n} = 0,$$

so by the Squeeze Theorem for Sequences,[45]

$$\lim_{n\to\infty} \Pr\left(\left|\overline{X}_n - \mathrm{E}[X]\right| \geq k\right) = 0. \ \square$$

[42] The equivalence follows from the complement rule (see Theorem 1.1.1): $\lim_{n\to\infty} \Pr(|X_n - c| < k) = \lim_{n\to\infty} [1 - \Pr(|X_n - c| \geq k)] = 1 - \lim_{n\to\infty} \Pr(|X_n - c| \geq k)$, so $\lim_{n\to\infty} \Pr(|X_n - c| \geq k) = 0 \iff \lim_{n\to\infty} \Pr(|X_n - c| < k) = 1$.

[43] A more general definition allows for X_n to converge in probability to a random variable X: for a sequence of random variables $(X_1, X_2, X_3, ...)$ and a random variable X, $X_n \xrightarrow{p} X$ if for any $k > 0$, $\lim_{n\to\infty} \Pr(|X_n - X| \geq k) = 0$.

[44] Technically we defined i.i.d. for a finite number of random variables, but the definition can easily be extended to infinitely many random variables.

[45] Squeeze Theorem for Sequences: if $\lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n = L$ and there exists an integer N such that $\forall n \geq N$, $b_n \leq a_n \leq c_n$, then $\lim_{n\to\infty} a_n = L$.

The WLLN is a profound result: as n gets large, the sample mean $\overline{X}$ becomes increasingly likely to approximate $\mathrm{E}[X]$ to any arbitrary degree of precision. We say that sample mean is *consistent* for the population mean. (We will formally define consistency in Section 2.1.5.)

As a shorthand, when we have a random variable that is implicitly a function of n i.i.d. draws, such as $\overline{X}$, we will henceforth typically dispense with the explicit invocation of sequences of random variables when discussing its asymptotic properties, i.e., its properties as the number of observations $n \to \infty$. So, for example, we will simply write $\overline{X} \xrightarrow{p} \mathrm{E}[X]$, without explicitly postulating the infinite sequence $(X_1, X_2, X_3, ...)$ and defining $\overline{X}_n$ to be the sample mean for the first n terms.

The WLLN has the following powerful implication.

Theorem 2.1.5. *Estimating the CDF*

Let X_1, X_2, ..., X_n be i.i.d. random variables with common CDF F. Let $x \in \mathbb{R}$ and let $Z_i = \mathrm{I}(X_i \leq x)$, where $\mathrm{I}(\cdot)$ is the indicator function, i.e., it takes the value 1 if the argument is true and 0 if it is false. Then

$$\overline{Z} \xrightarrow{p} F(x).$$

Proof: Let X_1, X_2, ..., X_n be i.i.d. random variables with common CDF F. Let $x \in \mathbb{R}$ and let $Z_n = \mathrm{I}(X_n \leq x)$. Then the PMF of Z is

$$f_Z(z) = \begin{cases} \Pr(X \leq x) & : z = 1 \\ \Pr(X > x) & : z = 0 \\ 0 & : \text{otherwise} \end{cases}$$

So, by the definition of the expected value,

$$\mathrm{E}[Z] = \sum_z z f_Z(z) = 1 \cdot \Pr(X \leq x) + 0 \cdot \Pr(X > x) = \Pr(X \leq x) = F(x).$$

Thus, by the Weak Law of Large Numbers,[46]

$$\overline{Z} \xrightarrow{p} \mathrm{E}[Z] = F(x). \ \square$$

We could apply this theorem for any or every value of x. Thus, we can always approximate the value of the CDF at any point to arbitrary precision. If n is very large, the probability that we are far off from the true value of $F(x)$ is very small.[47] So since the CDF tells us everything about a random variable, this means that we can estimate *any* feature of a random variable to arbitrary precision, given large enough n. (We show how this works in general in Section 2.1.5.)

[46] It is easy to show that Z_1, Z_2, ..., Z_n are i.i.d. random variables with finite $\mathrm{E}[Z]$ and finite $V(Z) > 0$.

[47] Stronger results are possible under weaker assumptions, namely the Strong Law of Large Numbers and the Glivenko-Cantelli Theorem. We will discuss the Glivenko-Cantelli Theorem briefly in Section 2.1.12.

2.1.2 Estimating the Sampling Variance

We now want to derive an estimator of $V(X)$, the *population variance* of a random variable X. (Again, this terminology will be explained in Section 2.1.3.) To do so, we will require a technical theorem, the *Continuous Mapping Theorem (CMT)*, which we state without proof.

Theorem 2.1.6. *The Continuous Mapping Theorem (CMT)*

Let $(X_1, X_2, X_3, ...)$ and $(Y_1, Y_2, Y_3, ...)$ be sequences of random variables and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a continuous function. If $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$, then $g(X_n, Y_n) \xrightarrow{p} g(a, b)$.

The CMT says that continuous functions preserve convergence in probability. Note that, though Theorem 2.1.6 states the CMT for the case of functions of two random variables, it also applies to functions of a single random variable and functions of more than two random variables.

We can now make our first attempt at proposing an estimator of $V(X)$. Recall from Theorem 1.4.5 that $V(X) = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2$. We know from the WLLN that the sample mean converges in probability to the expected value or population mean, so what if we simply "plug in" sample means for expected values? We will call this the *plug-in sample variance.*

Definition 2.1.4. *Plug-In Sample Variance*

For i.i.d. random variables X_1, X_2, ..., X_n, the plug-in sample variance is $\overline{X^2} - \overline{X}^2$.

Note that $\overline{X^2} = \sum_{i-1}^{n} X^2$, i.e., the sample mean of X^2.

The following theorem describes the key properties of the plug-in sample variance.

Theorem 2.1.7. *Properties of the Plug-In Sample Variance*

For i.i.d. random variables X_1, X_2, ..., X_n,

- $\mathrm{E}\left[\overline{X^2} - \overline{X}^2\right] = \dfrac{n-1}{n} V(X)$.
- $\overline{X^2} - \overline{X}^2 \xrightarrow{p} V(X)$.

Proof: Let X_1, X_2, ..., X_n be i.i.d. random variables and let $Z_i = X_i^2$, $\forall i \in \{1, 2, ..., n\}$. It follows from Theorem 1.3.4 that Z_1, Z_2, ..., Z_n are i.i.d., so by Theorem 2.1.1,

$$\mathrm{E}\left[\overline{X^2}\right] = \mathrm{E}\left[\overline{Z}\right] = \mathrm{E}\left[Z\right] = \mathrm{E}\left[X^2\right].$$

Now consider $\overline{X}^2$. Note that $\mathrm{E}\left[\overline{X}^2\right] \neq \mathrm{E}\left[X\right]^2$. Rather, since $V\left(\overline{X}\right) = \mathrm{E}\left[\overline{X^2}\right] - \mathrm{E}\left[\overline{X}\right]^2$,

$$\mathrm{E}\left[\overline{X}^2\right] = \mathrm{E}\left[\overline{X}\right]^2 + V\left(\overline{X}\right) = \mathrm{E}\left[X\right]^2 + \frac{V(X)}{n},$$

where the second equality follows from Theorem 2.1.1 and Theorem 2.1.2. Thus,

$$\begin{aligned}
\mathrm{E}\left[\overline{X^2} - \overline{X}^2\right] &= \mathrm{E}\left[\overline{X^2}\right] - \mathrm{E}\left[\overline{X}\right]^2 \\
&= \mathrm{E}\left[X^2\right] - \left(\mathrm{E}\left[X\right]^2 + \frac{V(X)}{n}\right) \\
&= \left(\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2\right) - \frac{V(X)}{n} \\
&= V(X) - \frac{V(X)}{n} \\
&= \left(1 - \frac{1}{n}\right) V(X) \\
&= \frac{n-1}{n} V(X).
\end{aligned}$$

Now, by the WLLN, $\overline{X^2} \xrightarrow{p} \mathrm{E}\left[X^2\right]$ and $\overline{X} \xrightarrow{p} \mathrm{E}\left[X\right]$. Let $g(u, v) = u - v^2$. Then by the CMT,

$$\overline{X^2} - \overline{X}^2 = g\left(\overline{X^2}, \overline{X}\right) \xrightarrow{p} g\left(\mathrm{E}\left[X^2\right], \mathrm{E}\left[X\right]\right) = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2 = V(X). \ \square$$

Theorem 2.1.7 says that the plug-in sample variance is biased but consistent. The size of the bias decreases as n increases (since $\frac{n-1}{n} \to 1$ as $n \to \infty$), so in large samples it becomes negligible. It is also easy to correct for this bias, while retaining the (more important) property of consistency. The *unbiased sample variance* (often referred to simply as the *sample variance*) incorporates this correction.

Definition 2.1.5. *Unbiased Sample Variance*

For i.i.d. random variables X_1, X_2, ..., X_n, the unbiased sample variance is

$$\hat{V}(X) = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right).$$

In general, the notation of adding a "hat" (i.e.,ˆ) will denote an estimated quantity. The following theorem states that the unbiased sample variance is indeed unbiased, as well as consistent.

Theorem 2.1.8. *Properties of the Unbiased Sample Variance*

For i.i.d. random variables X_1, X_2, ..., X_n,

- $\mathrm{E}\left[\hat{V}(X)\right] = V(X)$.
- $\hat{V}(X) \xrightarrow{p} V(X)$.

Proof: By linearity of expectations and Theorem 2.1.7,

$$\mathrm{E}\left[\hat{V}(X)\right] = \mathrm{E}\left[\frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right)\right] = \frac{n}{n-1}\mathrm{E}\left[\overline{X^2} - \overline{X}^2\right] = \frac{n}{n-1}\left(\frac{n-1}{n}V(X)\right) = V(X).$$

Now, by Theorem 2.1.7, $\overline{X^2} - \overline{X}^2 \xrightarrow{p} V(X)$. And treating $\frac{n}{n-1}$ as a degenerate random variable, it is easy to show that $\frac{n}{n-1} \xrightarrow{p} 1$. Thus, by the CMT,

$$\hat{V}(X) = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right) \xrightarrow{p} 1 \cdot V(X) = V(X). \ \square$$

So far, we have discussed the population variance $V(X)$, and an estimator thereof, the unbiased sample variance $\hat{V}(X)$. And in Section 2.1.1, we showed that the sampling variance of the sample mean is $V(\overline{X}) = {}^{V(X)}/_{n}$. (Do not confuse these three quantities!) We now want to propose an estimator of the sampling variance of the sample mean, $\hat{V}(\overline{X})$. The obvious choice is simply ${}^{\hat{V}(X)}/_{n}$.

Theorem 2.1.9. *Estimating the Sampling Variance*

For i.i.d. random variables X_1, X_2, ..., X_n, let $\hat{V}\left(\overline{X}\right) = \dfrac{\hat{V}(X)}{n}$. Then

- $\mathrm{E}\left[\hat{V}\left(\overline{X}\right)\right] = V\left(\overline{X}\right)$.
- $n\hat{V}\left(\overline{X}\right) \xrightarrow{p} nV\left(\overline{X}\right)$.

Proof: Let X_1, X_2, ..., X_n be i.i.d. random variables and let $\hat{V}\left(\overline{X}\right) = \dfrac{\hat{V}(X)}{n}$. Then

$$\mathrm{E}\left[\hat{V}\left(\overline{X}\right)\right] = \mathrm{E}\left[\frac{\hat{V}(X)}{n}\right] = \frac{\mathrm{E}\left[\hat{V}(X)\right]}{n} = \frac{V(X)}{n} = V\left(\overline{X}\right).$$

And by the CMT,

$$n\hat{V}\left(\overline{X}\right) = n\frac{\hat{V}(X)}{n} = \hat{V}(X) \xrightarrow{p} V(X) = nV\left(\overline{X}\right). \ \square$$

Thus, $\hat{V}(\overline{X})$ is an unbiased and consistent estimator of the sampling variance of the sample mean. It is therefore possible to estimate the average of a random variable (the population mean) and, furthermore, to estimate the uncertainty of that estimate. (We will formally discuss quantifying the uncertainty of an estimate in Section 2.2.)

As with any variance, we can take the square root of the sampling variance of the sample mean to obtain the standard deviation of the sampling distribution of the sample mean. This quantity is important enough to have its own name: the *standard error* of the sample mean.

Definition 2.1.6. *Standard Error of the Sample Mean*

For i.i.d. random variables X_1, X_2, ..., X_n, the standard error of the sample mean is

$$\sigma\left(\overline{X}\right) = \sqrt{V(\overline{X})}.$$

To estimate the standard error of the sample mean, we simply use $\hat{\sigma}(\overline{X}) = \sqrt{\hat{V}(\overline{X})}$.

When you see "standard error" in a paper, this is what the authors mean: an estimate thereof. We don't know the true standard error![48] This estimator is not unbiased,[49] but it is consistent.

Theorem 2.1.10. *Consistency of the Standard Error Estimator*

For i.i.d. random variables X_1, X_2, ..., X_n,

$$\sqrt{n}\,\hat{\sigma}\left(\overline{X}\right) \xrightarrow{p} \sqrt{n}\,\sigma\left(\overline{X}\right).$$

Proof: Let X_1, X_2, ..., X_n be i.i.d. random variables and let $g(u) = \sqrt{u}$. By Theorem 2.1.9, $n\hat{V}(\overline{X}) \xrightarrow{p} nV(\overline{X})$. So since g is continuous, by the CMT,

$$\sqrt{n}\,\hat{\sigma}\left(\overline{X}\right) = \sqrt{n\hat{V}\left(\overline{X}\right)} = g\left(n\hat{V}\left(\overline{X}\right)\right) \xrightarrow{p} g\left(nV\left(\overline{X}\right)\right) = \sqrt{nV\left(\overline{X}\right)} = \sqrt{n}\,\sigma\left(\overline{X}\right). \ \square$$

2.1.3 Random Sampling from a Population

So far, we have discussed how we can learn about the distribution of a random variable X (a mathematical construct) by observing repeated, independent realizations of that random variable, X_1, X_2, ..., X_n. We have been referring to the true properties of a random variable ($\mathrm{E}[X]$, $V(X)$, etc.) as "population" quantities. We now turn to the question suggested by this terminology: if X_1, X_2, ..., X_n represent units selected at random from some population, can we use these observations to estimate features of the population?

Suppose we have a *finite population* U consisting of N units, indexed $i = 1, 2, ...N$. Associated with each unit i is a response x_i. Define the *finite population mass function*[50] for U as:

$$f_{FP}(x) = \frac{1}{N}\sum_{i=1}^{N} I(x_i = x).$$

That is, $f_{FP}(x)$ is the proportion of units in U that have $x_i = x$. Let $\mathcal{X}$ denote the set of unique values of x_i, i.e., $\mathcal{X} = \{x \in \mathbb{R} : f_{FP}(x) > 0\}$. Let μ denote the population mean of U,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{x\in\mathcal{X}} x f_{FP}(x),$$

[48] Of course, since an estimator of the standard error is itself a random variable, it has its own sampling distribution and thus its own standard error, so we could, in theory, estimate the standard error of the standard error of the sample mean, thereby quantifying the uncertainty of our estimate of the uncertainty of our estimate of the population mean. However, since standard error estimators converge "quickly" to the true standard error, this is not a concern with large n.

[49] This is due to Jensen's Inequality: for a random variable X and a convex function $g : \mathbb{R} \to \mathbb{R}$, $\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$. Corollary: for a random variable X and a concave function $g : \mathbb{R} \to \mathbb{R}$, $\mathrm{E}[g(X)] \leq g(\mathrm{E}[X])$. These inequalities are strict when g is strictly convex/concave. The square root function is strictly concave, so

$$\mathrm{E}\left[\hat{\sigma}\left(\overline{X}\right)\right] = \mathrm{E}\left[\sqrt{\hat{V}\left(\overline{X}\right)}\right] < \sqrt{\mathrm{E}\left[\hat{V}\left(\overline{X}\right)\right]} = \sqrt{V\left(\overline{X}\right)} = \sigma\left(\overline{X}\right).$$

[50] This is not a standard term in statistics.

and let σ^2 denote the population variance of U,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \sum_{x\in\mathcal{X}}(x-\mu)^2 f_{FP}(x).$$

It will be helpful to have a concrete example to work with as we proceed.

Example 2.1.1. *Finite Population Random Sampling*

Suppose we have a finite population U consisting of $N = 4$ units. Let $x_1 = 3$, $x_2 = 4$, $x_3 = 3$, and $x_4 = 10$. Then:

- $\mathcal{X} = \{3, 4, 10\}$.
- $\mu = \frac{1}{4}(3 + 4 + 3 + 10) = 5$.
- $\sigma^2 = \frac{1}{4}(4 + 1 + 4 + 25) = \frac{17}{2}$.
- $f_{FP}(-1) = 0$.
- $f_{FP}(3) = \frac{1}{2}$.
- $f_{FP}(4) = \frac{1}{4}$.
- $f_{FP}(10) = \frac{1}{4}$.
- $f_{FP}(24603) = 0$.

Let's consider a random generative process that selects one unit from U at random (i.e., with all units having equal probability of being selected). Let the random variable X take on the value of x_i associated with the unit selected.[51] Each unit is selected with probability $1/4$, so we get $X = 3$ with probability $2/4 = 1/2$, $X = 4$ with probability $1/4$, and $X = 10$ with probability $1/4$. This is a probability mass function! Under random sampling, the distribution of outcomes in the population entirely determines the probability distribution of the random variable. The PMF of X *is* the finite population mass function: $f(x) = f_{FP}(x), \forall x \in \mathbb{R}$. And $\mathcal{X} = \text{Supp}\,(X)$. Thus,

- $\mathrm{E}\,[X] = \sum_x x f(x) = \sum_{x\in\mathcal{X}} x f_{FP}(x) = \mu = 5.$
- $V(X) = \mathrm{E}\,[(x - \mathrm{E}\,[X])] = \sum_x (x - \mathrm{E}\,[X])^2 f(x) = \sum_{x\in\mathcal{X}}(x-\mu)^2 f_{FP}(x) = \sigma^2 = \frac{17}{2}.$

So, under random sampling from U,

- $\mathrm{E}\,[X]$ is the *population mean*, i.e., the average of all of the x_i's.

[51] Formally, $\Omega = \{1, 2, 3, 4\}$; $P(\omega) = \frac{1}{4}, \forall\omega \in \Omega$; $X = \mathcal{X}(\omega)$; and $\mathcal{X}(\omega) = x_\omega, \forall\omega \in \Omega$.

- $V(X)$ is the *population variance*, i.e., the average squared deviation from the mean of all of the x_i's.

Nothing changes when we have *two* or more features associated with each unit. Suppose we had a double (x_i, y_i) associated with each unit i. When we randomly draw unit i, we observe and record both values in the double. Then we would just have a joint PMF for the random vector (X, Y), and we could estimate covariances, conditional expectations, conditional PMFs, etc. Each of these have population analogues (population covariances, conditional means, etc.).

There's a catch, though. Remember that we have talked about a sample mean as being the average of a collection of n i.i.d. draws of a random variable. The i.i.d. model excludes *without replacement* sampling when N is finite. Why? When we draw units from a finite population without replacement, the distribution of the second draw X_2 depends on the first draw X_1.

Recall from Theorem 1.2.9 that X_1 and X_2 are independent if and only if, $\forall (x_1, x_2) \in \mathbb{R}^2$ with $f_{X_1}(x_1) > 0$, $f_{X_2|X_1}(x_2|x_1) = f_{X_2}(x_2)$. In Example 2.1.1, $f_{X_2}(10) = \Pr(X_2 = 10) = 1/4$, but $f_{X_2|X_1}(10|10) = \Pr(X_2 = 10|X_1 = 10) = 0$. Therefore, without replacement sampling does not yield i.i.d. samples. Processes that do yield i.i.d. draws include:

- with replacement sampling,
- without replacement sampling when $N \gg n$, and
- other types of random processes.

In the case of with replacement sampling, we "put back" each unit after we draw it, so the population distribution is not changed by the act of taking a draw. Thus, every draw is taken from the same distribution, independently. In the case of without replacement sampling from a very large population, removing a single unit or a few units changes the distribution of the remaining units by a negligible amount. We can therefore effectively treat the population as infinite. For example, suppose we had a population of $1,000,000$ people consisting of exactly $500,000$ women and $500,000$ men. Suppose we sample n people without replacement. Let $X_i = 0$ if person i in our sample is a woman and $X_i = 1$ if person i is a man. Then $\Pr(X_2 = 1) = 500,000/1,000,000 = 0.5$, and $\Pr(X_2 = 1|X_1 = 1) = 499,999/999,999 \approx 0.5$. Similarly, the probability that person n is a woman will be approximately 0.5 regardless of the genders of persons 1 through $n-1$, so long as n is a very small fraction of $1,000,000$. Finally, other types of random process can yield i.i.d. outcomes, such as coin flips, die rolls, etc.

2.1.4 The Central Limit Theorem

We now present what is perhaps the most profound and important theorem in statistics, the *Central Limit Theorem (CLT)*. Before we can state this theorem, we require a couple of definitions.

Definition 2.1.7. *Convergence in Distribution*

Let $(X_1, X_2, X_3, ...)$ be a sequence of random variables with CDFs $(F_1, F_2, F_3, ...)$, and let X be a random variable with CDF F. Then X_n converges in distribution to X if, $\forall x \in \mathbb{R}$ at which F is continuous,

$$\lim_{n\to\infty} F_n(x) = F(x),$$

We write $X_n \xrightarrow{d} X$ to denote that X_n converges in distribution to X.

In other words, X_n converges in distribution to X if as $n \to \infty$ the CDF of X_n approaches the CDF of X at every point.

We must also define the *standardized sample mean*.

Definition 2.1.8. *Standardized Sample Mean*

For i.i.d. random variables X_1, X_2, ..., X_n with finite $\mathrm{E}[X] = \mu$ and finite $V(X) = \sigma^2 > 0$, the standardized sample mean is:

$$Z = \frac{\left(\overline{X} - \mathrm{E}\left[\overline{X}\right]\right)}{\sigma\left(\overline{X}\right)} = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma}.$$

It is easy to show that $\mathrm{E}[Z] = 0$ and $V(Z) = \sigma(Z) = 1$. These "nice" properties are the reason why Z is called the standardized sample mean.

We can now state the Central Limit Theorem for the sample mean.

Theorem 2.1.11. *The Central Limit Theorem (CLT)*

Let X_1, X_2, ..., X_n be i.i.d. random variables with finite $\mathrm{E}[X] = \mu$ and finite $V(X) = \sigma^2 > 0$, and let Z be the standardized sample mean. Then

$$Z \xrightarrow{d} N(0, 1),$$

or equivalently,[52]

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} N(0, \sigma^2).$$

We omit the proof of this theorem.[53] In words, the CLT says that, if n is large, the sampling distribution of the standardized sample mean is approximately standard normal. This is extremely useful because it will allow us to rigorously quantify our uncertainty when n is large, as we will see in Section 2.2.

2.1.5 Estimation Theory

Thus far, we have spoken somewhat loosely about estimators and their properties: bias, consistency, and so forth. We now define these concepts in more formal and general terms. Suppose there is some feature

[52] The equivalence follows from the CMT, which, though we stated it for convergence in probability, also holds for convergence in distribution.

[53] Full proofs of the CLT invariably require fairly advanced mathematics, though the proof of a special case, the *De Moivre-Laplace Theorem*, is relatively straightforward.

θ associated with random variable X. We observe a sample of n i.i.d. draws of X, denoted by the vector $\mathbf{x} = (x_1, x_2, ..., x_n)^T$. Thus, $\mathbf{x}$ is a single draw of the random vector $\mathbf{X} = (X_1, X_2, ..., X_n)^T$. Using this sample, we derive an *estimate* of θ using some function $h(\mathbf{x})$. Let $t = h(\mathbf{x}) = h(x_1, x_2, ..., x_n)$ denote the estimate. The *estimator* is the random variable $\hat{\theta} = h(\mathbf{X}) = h(X_1, X_2, ..., X_n)$ that takes on the value of the estimate.

We can now formally define some estimator properties. A given estimator may or may not satisfy any one of these properties.

Definition 2.1.9. *Unbiasedness*

An estimator $\hat{\theta}$ is unbiased for θ if $\mathrm{E}[\hat{\theta}] = \theta$.

An unbiased estimator gets the right answer "on average." This is a good property for an estimator to have, though it is (in our view) less important than it might seem.

Definition 2.1.10. *Consistency*

An estimator T is consistent for θ if $\hat{\theta} \xrightarrow{p} \theta$.

The WLLN can thus be stated as: "The sample mean is consistent for the population mean." We typically say that a sampling variance estimator $\hat{V}(\hat{\theta})$ is consistent if and only if a stronger statement holds, namely $n\hat{V}(\hat{\theta}) \xrightarrow{p} nV(\hat{\theta})$.[54] If $\hat{\theta}$ is consistent for θ, then $\lim_{n\to\infty} V(\hat{\theta}) = 0$.[55] Therefore, the fact that $\hat{V}(\hat{\theta}) \xrightarrow{p} V(\hat{\theta})$ doesn't really say anything meaningful about how well $\hat{V}(\hat{\theta})$ approximates $V(\hat{\theta})$ in finite samples, since *any* $\hat{V}(\hat{\theta})$ such that $\hat{V}(\hat{\theta}) \xrightarrow{p} 0$, e.g., $\hat{V}(\hat{\theta}) = 1/n^3$, would satisfy this property. Similarly, we say that a standard error estimator $\hat{\sigma}(\hat{\theta})$ is consistent if $\sqrt{n}\,\hat{\sigma}(\hat{\theta}) \xrightarrow{p} \sqrt{n}\,\sigma(\hat{\theta})$.

Definition 2.1.11. *Asymptotic Normality*

An estimator $\hat{\theta}$ is asymptotically normal if $\hat{\theta} \xrightarrow{d} N(\theta, \phi^2/n)$*, for some finite* $\phi > 0$.

The CLT can thus be stated as: "The sample mean is asymptotically normal." Again, not all estimators satisfy all (or any) of these properties. For example:

- The plug-in sample variance is not unbiased for the population variance, but it is consistent.
- The unbiased sample variance is both unbiased and consistent for the population variance.
- $\hat{\theta} = h(X_1, ..., X_n) = 3$ does not generally satisfy any of the above properties.
- $\hat{\theta} = h(X_1, ..., X_n) = X_1$ is unbiased for the sample mean, but it is not generally consistent.

[54] There is also a stronger notion of consistency for an estimator: root-n consistency, which implies that $\hat{\theta} - \theta$ is of the same order of magnitude as $1/\sqrt{n}$. This is important for establishing consistency of a sampling variance estimator under the above definition, since otherwise $nV(\hat{\theta})$ may grow infinitely large as n grows.

[55] This follows from Chebyshev's Inequality.

2.1.6 The Plug-In Principle[56]

To avoid having to consider the discrete and continuous cases separately, we shall use the following notation in this section. Let F be a CDF of a random variable X, and let f be the corresponding PMF/PDF. Then let

$$\int g(x)dF(x) = \begin{cases} \sum_x g(x)f(x) & : F \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & : F \text{ is continuous} \end{cases}$$

The estimation philosophy used in this chapter has been informally based on the *plug-in principle*: write down the feature of the population that you're interested in, and then use the sample analogue to estimate it. E.g., if you want to estimate the expected value, use the sample mean. If you want to estimate the population variance, use the sample variance. It is possible to formalize and generalize this idea. In general, the CDF of a random variable tells us *everything* about the random variable. The CDF of X is

$$F(x) = \Pr(X \leq x) = \mathrm{E}\left[\mathrm{I}\left(X \leq x\right)\right].$$

There is a sample analogue to the CDF, the *empirical CDF (ECDF)*.

Definition 2.1.12. *Empirical CDF (ECDF)*

For i.i.d. random variables X_1, X_2, ..., X_n, the ECDF $\hat{F}$ of X is

$$\hat{F}(x) = \overline{\mathrm{I}\left(X \leq x\right)}, \forall x \in \mathbb{R}.^{57}$$

Just as the CDF is a function that fully describes the population, the ECDF is a function that fully describes the sample. Note that the ECDF is always discrete. There are two (equivalent) ways of thinking about how $\hat{F}(x)$ is computed:

- replacing the expected value in the definition of the CDF with a sample mean.[58]
- counting the number of observations i satisfying $X_i \leq x$ and dividing by n.

We know, because of the properties of the sample mean, that, $\forall x \in \mathbb{R}$, $\hat{F}(x)$ will be unbiased and consistent for $F(x)$, (see Theorem 2.1.5).[59]

[56] This section borrows heavily from Wasserman (2004), Wasserman (2005), and Wasserman (2012).

[57] This definition, and everything that follows, can easily be generalized to the multivariate case. E.g., in the bivariate case, for i.i.d. random vectors (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), the *joint* ECDF of (X, Y) is $\hat{F}(x, y) = \overline{\mathrm{I}\left(X \leq x, Y \leq y\right)}$. Everything else proceeds analogously.

[58] Recall from Definition 2.1.2 that $\overline{\mathrm{I}\left(X \leq x\right)}$ denotes the sample mean of $\mathrm{I}\left(X \leq x\right)$.

[59] In fact, an even stronger claim holds. The Glivenko-Cantelli Theorem:

$$\sup_x |\hat{F}(x) - F(x)| \overset{a.s.}{\to} 0.$$

The notation $\overset{a.s.}{\to}$ denotes *almost sure convergence*, a stronger notion of convergence than either convergence in probability or convergence in distribution. (We will not discuss this type of convergence in detail.) In words, the Glivenko-Cantelli Theorem states that the biggest difference between the CDF and ECDF converges to zero. In fact, it converges to zero "quickly." See the Dvoretzky-Kiefer-Wolfowitz inequality.

We now introduce the concept of a *statistical functional*. Simply put, a statistical functional is a function of the CDF.

Definition 2.1.13. *Given a set of CDFs $\mathcal{F}$, a statistical functional is a function $T : \mathcal{F} \to \mathbb{R}$.*

E.g., for the random variable X with CDF F, we can write the expected value of X as

$$\mathrm{E}[X] = T_E(F) = \int x dF(x).$$

Likewise, we can write the variance of X as

$$V[X] = T_V(F) = \int (x - \mathrm{E}[X])^2 dF(x) = \int x^2 dF(x) - \mathrm{E}[X]^2 = \int x^2 dF(x) + \left[\int x dF(x)\right]^2.$$

Consider θ that can be written as $T(F)$. (This is remarkably general.) We can define the *plug-in estimator* of θ as follows.

Definition 2.1.14. *Plug-In Estimator*

For i.i.d. random variables X_1, X_2, ..., X_n with common CDF F, the plug-in estimator of $\theta = T(F)$ is

$$\hat{\theta} = T(\hat{F}).$$

To estimate $\theta = T(F)$, simply apply the same functional to the *empirical CDF*. For example, since the expected value of X is $\mathrm{E}[X] = T_E(X) = \int x dF(x)$, the plug-in estimator of $\mathrm{E}[X]$ is

$$\begin{aligned}
\hat{\mathrm{E}}[X] &= T_E(\hat{F}) \\
&= \int x d\hat{F}(x) \\
&= \sum_x x \hat{f}(x) \\
&= \sum_x x \frac{(\#\text{ of observations } i \text{ with outcome } X_i = x)}{n} \\
&= \frac{1}{n} \sum_x x (\#\text{ of observations } i \text{ with outcome } X_i = x) \\
&= \frac{1}{n} \sum_{i=1}^{n} x = \overline{X},
\end{aligned}$$

where $\hat{f}$ is the empirical PMF. Thus, the sample mean is the plug-in estimator of the expected value. Similarly, since the variance of X is $V[X] = T_V(F) = \int x^2 dF(x) + [\int x dF(x)]^2$, the plug-in estimator of $V(X)$ is

$$\hat{V}(X) = T_V(\hat{F}) = \int x^2 d\hat{F}(x) - \left[\int x d\hat{F}(x)\right]^2 = \overline{X^2} - \overline{X}^2.$$

Thus, the plug-in sample variance is, in fact, the plug-in estimator of the variance.

Typically we don't need to go to the extreme of writing everything in terms of the CDF or ECDF, as we can derive most plug-in estimators simply by substituting sample means for expected values. Nonetheless, when we talk about a plug-in estimator, we're fundamentally talking about an estimator that substitutes ("plugs in") the ECDF for the CDF. Plug-in estimators are usually* consistent, but not generally unbiased.[60]

The idea of a plug-in estimator is quite profound. We want to know a feature θ of the CDF. We express θ as a function T of the CDF. We then observe the sample analogue of the CDF, the ECDF. We know that, as n grows large, the ECDF tends to more and more closely approximate the CDF. We therefore estimate θ by applying the function T to the ECDF. Then, as n grows large, the estimate θ tends to more and more closely approximate θ.

2.1.7 Kernel Estimation

The procedure illustrated above works well for discrete random variables. But what if we are dealing with a continuous random variable? If we are willing to impose some strong assumptions about the distribution of the random variable, it becomes fairly easy to estimate its PDF (see Chapter 7). But what if we don't want to make such strong assumptions? In that case, we can still estimate the PDF using *kernel density estimation*.

The key idea behind kernel density estimation is to "smooth" the data so as to get estimates of the PDF everywhere. We can estimate the value of the PDF at a given point x by calculating the proportion of the observations that have X_i near x. It's actually quite intuitive.

Definition 2.1.15. *Kernel Density Estimator*

Let X_1, X_2, ..., X_n be i.i.d. continuous random variables with common PDF f. Let $K : \mathbb{R} \to \mathbb{R}$ be a symmetric function satisfying $\int_{-\infty}^{\infty} K(x)dx = 1$, and let $K_h(x) = \frac{1}{h}K(\frac{x}{h})$, $\forall x \in \mathbb{R}$ and $h > 0$. Then a kernel density estimator of $f(x)$ is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \forall x \in \mathbb{R}.$$

The function K is called the kernel and the scaling parameter h is called the bandwidth.

The kernel and bandwidth tell us how much to weight each observation X_i depending on its distance from the point of interest x. The farther an observation is from the point of interest, the less it counts. This is easiest to understand in the univariate case, though everything carries through to multiple variables so long as "distance" is defined accordingly. (Regular Euclidean distance is the usual choice.) The kernel tells us the shape of the weighting, and the bandwidth tells us how far out we should go, i.e., the width or scale of the kernel. There are many common kernels (see Figure 2.1.1). Then computing estimates of the densities at every point is straightforward: we just add up all of the observations, weighted by a metric of how much we want each one to count. The closer it is to the point of interest, the more it counts.

[60] By "usually*" we mean probably safe to assume in practice, but don't blame us if you encounter some weird exception.

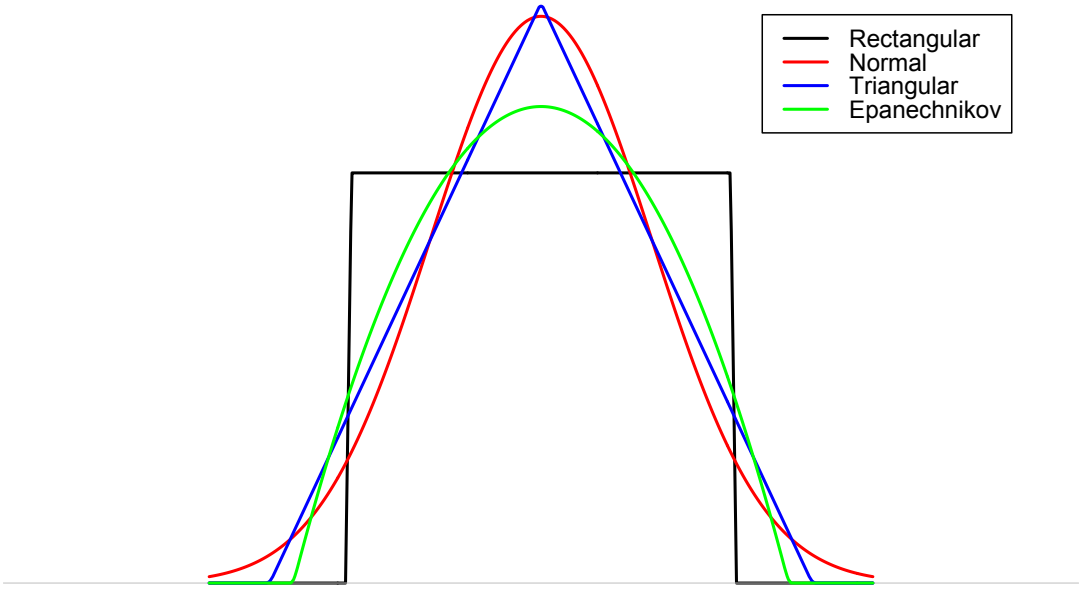


Figure 2.1.1. Visualization of different kernels.

How do we ensure that our kernel density estimator is consistent? Simply let the bandwidth $h \to 0$ as $n \to \infty$. Then as long as the PDF really is continuous, kernel density estimation will be consistent.[61] Why does this work? Because of the definition of the PDF!

$$f(x) = \frac{dF(x)}{dx} = \lim_{h \to 0^+} \frac{F(x+h) - F(x-h)}{2h}.$$

Consider the case of a rectangular kernel. For every $x \in \mathbb{R}$, the ECDF $\hat{F}(x)$ is consistent for the CDF

[61] We also want to make sure that $h \to 0$ at an appropriate rate, so that we're never over- or undersmoothing. Silverman's "rule of thumb" (Silverman 1986, page 48, equation 3.31) for kernel density estimation with a normal (Gaussian) kernel is to let

$$h = \frac{0.9 \min\left(\hat{\sigma}(X), \widehat{IQR}/1.34\right)}{n^{1/5}},$$

where $\hat{\sigma}(X) = \sqrt{\hat{V}(X)}$ and $\widehat{IQR}$ is the (sample) interquartile range. Under this rule, the range of observations that we count toward a point's density is going to shrink at a rate on the order of $n^{-1/5}$. That's pretty slow, but eventually, with large enough n, we'll only count observations that are very close.

$F(x)$ by the WLLN (see Theorem 2.1.5), so by the CMT, the kernel density estimate

$$\begin{aligned}
\hat{f}(x) &= \lim_{h \to 0^+} \frac{\frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\left(x - h \leq X_i \leq x + h\right)}{2h} \\
&= \lim_{h \to 0^+} \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h} \\
&\xrightarrow{p} \lim_{h \to 0^+} \frac{F(x+h) - F(x-h)}{2h} \\
&= \frac{dF(x)}{dx} \\
&= f(x).
\end{aligned}$$

Now, once you have an empirical estimate of the PDF, you can compute any quantity that you want to off of the estimate. You can integrate your estimate of the PDF to obtain an estimate of the full CDF, and thus you can apply any statistical functional to that estimate. The main benefit of this approach comes in the multivariate case, since by obtaining a continuous estimate of the joint PDF, you can get point predictions very straightforwardly even when you need to condition on a continuous variable.

For example, if we used kernel density estimation to estimate $f_{Y|X}(y|\mathbf{x})$ (simply use kernel estimation to estimate $f(y, \mathbf{x})$ and $f_{\mathbf{X}}(\mathbf{x})$ and then apply Definition 1.3.9), then we could use our estimate $\hat{f}_{Y|\mathbf{X}}(y|\mathbf{x})$ to estimate the CEF at any given point $\mathbf{x}$:

$$\hat{\mathrm{E}}\left[Y|\mathbf{X} = \mathbf{x}\right] = \int_{-\infty}^{\infty} y \hat{f}_{Y|\mathbf{X}}(y|\mathbf{x}) dy.$$

So long as the density estimator is consistent, all of the associated plug-in estimators will be consistent. This holds for any reasonable feature of the random variable that you might be interested in (moments, conditional variances, etc.).

2.2 Inference

We've shown how to estimate features of a random variable or population from i.i.d. samples. But how do we know how certain we can be that these estimates closely approximate reality? Can we precisely quantify the degree of uncertainty surrounding our estimates? These questions are the subject of statistical *inference*. In this section, among other things, we shall see the importance of our discussion of the standard error and the Central Limit Theorem in Section 2.1.

2.2.1 Confidence Intervals

A *confidence interval* for θ, loosely speaking, is an *interval estimate* that covers the true value of θ with at least a given probability. In other words, if we were to repeatedly take i.i.d. samples of size n of a random

variable X and compute a (valid) 95% confidence interval for some θ from each of those samples, then 95 percent of the time, the interval should include θ. The percentage of the time that the confidence interval covers θ is known as the *coverage* rate. The following theorem gives a very general method for obtaining asymptotically valid confidence intervals.

Theorem 2.2.1. *Normal Approximation-Based Confidence Interval*

Let $\hat{\theta}$ be an asymptotically normal estimator of θ, $n\hat{V}(\hat{\theta})$ be a consistent estimator of $nV(\hat{\theta})$, and $\alpha \in (0,1)$. Then an asymptotically valid "Wald-type" normal approximation-based confidence interval for θ with coverage $(1-\alpha)$ is given by

$$\widehat{CI}_{1-\alpha}(\theta) = \left(\hat{\theta} - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})}, \hat{\theta} + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})}\right)$$

where z_c denotes the c^{th} quantile of the standard normal distribution, i.e., $\Phi(z_c) = c$, $\forall c \in (0,1)$. For any given $\alpha \in (0,1)$, we say that the confidence level *is $100(1-\alpha)\%$ and that $\widehat{CI}_{1-\alpha}(\theta)$ is a $100(1-\alpha)\%$ confidence interval for θ.*

By "asymptotically valid,"[62] we mean that

$$\lim_{n\to\infty} \Pr\left(\theta \in \widehat{CI}_{1-\alpha}(\theta)\right) \geq 1-\alpha.$$

Proof: Suppose $\hat{\theta}$ is an asymptotically normal estimator of θ and $n\hat{V}(\hat{\theta})$ is a consistent estimator of $nV(\hat{\theta})$.

Define the random variable

$$Z = \frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}}.$$

Then by asymptotic normality and the CMT,[63] $Z \xrightarrow{d} N(0,1)$. Now we plug in the *estimated* standard error. Define the random variable

$$Z' = \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}} = \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}} \frac{\sqrt{n}\sqrt{V(\hat{\theta})}}{\sqrt{n}\sqrt{V(\hat{\theta})}} = \frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}} \frac{\sqrt{nV(\hat{\theta})}}{\sqrt{n\hat{V}(\hat{\theta})}} = Z\frac{\sqrt{nV(\hat{\theta})}}{\sqrt{n\hat{V}(\hat{\theta})}}.$$

Then by the CMT and consistency of the variance estimator,

$$\frac{\sqrt{nV(\hat{\theta})}}{\sqrt{n\hat{V}(\hat{\theta})}} \xrightarrow{p} 1.$$

[62] There is a technical distinction between *pointwise* and *uniform* asymptotic validity; with the latter being a stronger notion. Here we consider pointwise asymptotic validity.

[63] Again, we stated the CMT for convergence in probability, but it also holds for convergence in distribution.

So by Slutsky's Theorem,[64] $Z' \xrightarrow{d} N(0,1)$. Now, take an arbitrary $\alpha \in (0,1)$. Note that by symmetry of the normal distribution, $z_{\alpha/2} = -z_{1-\alpha/2}$. Since $Z' \xrightarrow{d} N(0,1)$, it follows that

$$\begin{aligned}
\lim_{n\to\infty} \Pr\left(-z_{1-\alpha/2} \leq Z' \leq z_{1-\alpha/2}\right) &= \lim_{n\to\infty} \left[F_{Z'}\left(z_{1-\alpha/2}\right) - F_{Z'}\left(-z_{1-\alpha/2}\right)\right] \\
&= \lim_{n\to\infty} \left[F_{Z'}\left(z_{1-\alpha/2}\right) - F_{Z'}\left(z_{\alpha/2}\right)\right] \\
&= \lim_{n\to\infty} F_{Z'}\left(z_{1-\alpha/2}\right) - \lim_{n\to\infty} F_{Z'}\left(z_{\alpha/2}\right) \\
&= \Phi\left(z_{1-\alpha/2}\right) - \Phi\left(z_{\alpha/2}\right) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,
\end{aligned}$$

Where $F_{Z'}$ denotes the CDF of Z'. Thus,

$$\begin{aligned}
\lim_{n\to\infty} \Pr\left(\theta \in \widehat{CI}_{1-\alpha}(\theta)\right) &= \lim_{n\to\infty} \Pr\left(\hat{\theta} - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})} \leq \theta \leq \hat{\theta} + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})}\right) \\
&= \lim_{n\to\infty} \Pr\left(-z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})} \leq \theta - \hat{\theta} \leq z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})}\right) \\
&= \lim_{n\to\infty} \Pr\left(-z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})} < \hat{\theta} - \theta < z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})}\right) \\
&= \lim_{n\to\infty} \Pr\left(-z_{1-\alpha/2} < \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}} < z_{1-\alpha/2}\right) \\
&= \lim_{n\to\infty} \Pr\left(-z_{1-\alpha/2} < Z' < z_{1-\alpha/2}\right) \\
&= 1 - \alpha. \ \square
\end{aligned}$$

Specific values of $z_{1-\alpha/2}$ can be obtained using a Z table, graphing calculator, or statistical software, but it is useful to know the most commonly used values:

- For a 90% confidence interval, $\alpha = 0.10$ and $z_{1-\alpha/2} \approx 1.65$.
- For a 95% confidence interval, $\alpha = 0.05$ and $z_{1-\alpha/2} \approx 1.96$.

[64] Slutsky's Theorem: let $(X_1, X_2, X_3, ...)$ and $(Y_1, Y_2, Y_3, ...)$ be sequences of random variables. Let X be a random variable and $c \in \mathbb{R}$. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

- $X_n + Y_n \xrightarrow{d} X + c$.
- $X_n Y_n \xrightarrow{d} cX$.
- $X_n/Y_n \xrightarrow{d} X/c$, provided that $c \neq 0$.

- For a 99% confidence interval, $\alpha = 0.01$ and $z_{1-\alpha/2} \approx 2.58$.

Example 2.2.1. *Confidence Interval for the Population Mean*

Suppose we had n i.i.d. draws of a random variable X. Then an approximate 95% confidence interval for the population mean $\mu = \mathrm{E}[X]$ would be

$$\begin{aligned}
\widehat{CI}_{.95}(\mu) &= \left(\overline{X} - 1.96\sqrt{\frac{\hat{V}(X)}{n}} \, , \overline{X} + 1.96\sqrt{\frac{\hat{V}(X)}{n}} \right) \\
&= \left(\overline{X} - 1.96\sqrt{\hat{V}\left(\overline{X}\right)} \, , \overline{X} + 1.96\sqrt{\hat{V}\left(\overline{X}\right)} \right) \\
&= \left(\overline{X} - 1.96\,\hat{\sigma}\left(\overline{X}\right) \, , \overline{X} + 1.96\,\hat{\sigma}\left(\overline{X}\right) \right).
\end{aligned}$$

If n is large, then among many i.i.d. samples of size n, 95% of the time $\widehat{CI}_{.95}(\mu)$ will encompass the true value of μ.

2.2.2 Hypothesis Testing

The foregoing discussion of confidence intervals naturally leads to the closely related idea of hypothesis testing. Suppose we wanted to test the *null hypothesis* that $\theta = \theta_0$. That is, having observed our sample and computed an estimate $\hat{\theta}$, we could ask: if θ were actually equal to θ_0, what would be the probability that we would have obtained a $\hat{\theta}$ at least as far off from θ_0 as we did? This probability is called a *p-value*.[65]

Definition 2.2.1. *p-Value*

Let $\hat{\theta}$ be an estimator of θ, and let $\hat{\theta}^*$ be the observed value of $\hat{\theta}$. Then for the null hypothesis that $\theta = \theta_0$,

- *a lower one-tailed p-value for $\hat{\theta}^*$ is given by $p = \Pr\left(\hat{\theta} \leq \hat{\theta}^*\right)$,*
- *a upper one-tailed p-value for $\hat{\theta}^*$ is given by $p = \Pr\left(\hat{\theta} \geq \hat{\theta}^*\right)$,*
- *a two-tailed p-value for $\hat{\theta}^*$ is given by $p = \Pr\left(|\hat{\theta} - \theta_0| \geq |\hat{\theta}^* - \theta_0|\right)$,*

where the probabilities are calculated under the assumption that $\theta = \theta_0$.

The following theorem gives a general method for obtaining asymptotically valid p-values using the same principles as Theorem 2.2.1.

[65] This is only one way of computing a p-value. It is possible to test hypotheses about θ using any variety of *test statistics*; we focus here on the case of using $\hat{\theta}$ as a test statistic.

Theorem 2.2.2. *Normal Approximation-Based p-Values*

Let $\hat{\theta}$ be an asymptotically normal estimator of θ, $n\hat{V}(\hat{\theta})$ be a consistent estimator of $nV(\hat{\theta})$. Let $\hat{\theta}^*$ be the observed value of $\hat{\theta}$. Then

- *an asymptotically valid lower one-tailed p-value for $\hat{\theta}^*$ is given by*

$$p = \Phi\left(\frac{\hat{\theta}^* - \theta_0}{\sqrt{\hat{V}(\hat{\theta})}}\right).$$

- *an asymptotically valid upper one-tailed p-value for $\hat{\theta}^*$ is given by*

$$p = 1 - \Phi\left(\frac{\hat{\theta}^* - \theta_0}{\sqrt{\hat{V}(\hat{\theta})}}\right).$$

- *an asymptotically valid two-tailed p-value for $\hat{\theta}^*$ is given by*

$$p = 2\left(1 - \Phi\left(\frac{|\hat{\theta}^* - \theta_0|}{\sqrt{\hat{V}(\hat{\theta})}}\right)\right).$$

We omit the proof of this theorem, as well as a formal definition of asymptotic validity. Loosely speaking, however, asymptotic validity in the context of p-values implies that, under the null hypothesis that $\theta = \theta_0$, as $n \to \infty$, $p \xrightarrow{d} U(0,1)$. (E.g., if $\theta = \theta_0$, n is large and the p-value is asymptotically valid, then across repeated samples, we would only obtain $p \leq 0.05$ in 5% of samples.)

The intuition behind Theorem 2.2.2 should be clear: if n is large, then under the assumption that $\theta = \theta_0$, $\hat{\theta} \overset{approx.}{\sim} N(\theta_0, V(\hat{\theta}))$ (by asymptotic normality), and $\hat{V}(\hat{\theta}) \approx V(\hat{\theta})$ (by consistency), so $\hat{\theta} \overset{approx.}{\sim} N(\theta_0, \hat{V}(\hat{\theta}))$. Thus, we can estimate, e.g., $\Pr(\hat{\theta} \leq \hat{\theta}^*)$ using the normal CDF.

Traditionally, we *reject* the null hypothesis that $\theta = \theta_0$ if we obtain a p-value that is below some conventional significance threshold (usually .05) and *fail to reject* (note: not "accept") the null hypothesis otherwise. Be careful, though: lots of tests means lots of "significant" results.

2.2.3 The Bootstrap

The idea of using the ECDF to approximate a CDF motivates our final topic for this section: the bootstrap. The bootstrap provides a procedure for estimating standard errors for *any* statistic of interest. This procedure is usually* consistent. The bootstrap derives its name from the phrase "pull yourself up by your own bootstraps." This is an apt description, as we shall see shortly.

If we *knew* the CDF of some random variable, then we'd be able to know what the sampling distribution of any sample statistic would be. E.g., if we knew that we had a fair coin, we'd know that the sampling distribution of the sample mean would have mean $1/2$ and variance $1/4n$. In fact, we could use computer

simulation to calculate the sampling distribution of the sample mean to arbitrary precision by taking repeated simulated samples of size n, calculating the sample mean of each simulated sample, and observing the distribution of the resulting values.

So characterizing the sampling variability of the sample mean (for a given n) is straightforward *if* we know the distribution that we're sampling from. The inferential problem, however, is that we *don't* know the distribution that we're sampling from. The bootstrap solution to this problem is quite elegant: simply pretend that the distribution that you're sampling from looks *exactly* like the sample that you have. I.e., *plug-in* the ECDF for the CDF and sample from *that*. It's a *resampling method*.

More specifically, to characterize the uncertainty of your estimator:

1. Take a *with replacement* sample of size n *from your sample*.
2. Calculate your would-be estimate using this *bootstrap sample*.
3. Repeat steps 1 and 2 many, many times.
4. Using the resulting collection of *bootstrap estimates*, calculate the standard deviation of the *bootstrap distribution* of your estimator.

This standard deviation will usually consistently estimate the true standard error. If the asymptotic distribution of the estimator is normal (which it usually* is), you can construct confidence intervals and p-values in the usual way.[66]

Why does this work? Because the bootstrap is a plug-in estimator! Under the i.i.d. sampling framework, the sampling distribution of an estimator is determined entirely by the CDF and n. Since the ECDF looks more and more like the CDF as n gets large, the plug-in sampling distribution yielded by the bootstrap looks more and more like the true sampling distribution.

The bootstrap is particularly useful in cases where you want to conduct inference using a weird estimator, e.g., $(\hat{\beta}_0 - \hat{\beta}_1)^3/(2 + \hat{\beta}_2^2)$. Although an analytic variance estimator might be possible to derive, it might be difficult, or it might require using asymptotic approximations that yield very bad finite-sample performance. So whenever you have i.i.d. sampling, the bootstrap is usually a sensible way to estimate standard errors, and it saves you the hassle of having to work out an analytic variance estimator in cases where it might be difficult. For more details on the bootstrap, we suggest Efron and Tibshirani (1994).

2.3 Cluster Samples

How does clustering in our sampling affect our ability to make inferences? Thus far we have been considering i.i.d. observations of a single-valued random variable. Let us now assume that we have m i.i.d.

[66] There are alternative ways to construct confidence intervals (and p-values) with the bootstrap, including the quantile approach. Assuming that we have B bootstrap replicates, the quantile bootstrap estimate of a 95% confidence interval can be computed by sorting the B bootstrap estimates from smallest to largest, and forming an interval by taking the $\lfloor 0.025 \times B \rfloor$th sorted bootstrap estimate and the $\lceil 0.975 \times B \rceil$th sorted bootstrap estimate. The quantile approach often has better performance when n is small.

clusters of observations, each consisting of k values and represented by a random vector $(X_{i1}, X_{i2}, ...X_{ik})$. Thus, while observations must be independent across clusters (e.g., $X_{ij} \perp\!\!\!\perp X_{i'j}$), within clusters they may be statistically dependent (i.e., it is not generally the case that $X_{ij} \perp\!\!\!\perp X_{ij'}$). Let $n = mk$, so that the number of observations is the number of clusters times the number of observations per cluster. What can we learn from such a sample?

2.3.1 Estimation with Clustering

In general, the random vector $(X_{i1}, X_{i2}, ...X_{ik})$ is characterized by a joint CDF F. Taking m i.i.d. draws of this random vector yields an empirical joint CDF $\hat{F}$. Statistical functionals are now simply written as functions of the joint CDF and estimated using the empirical joint CDF.

Let's see how this works for the sample mean. It will be useful for us to define the following random variable, the *cluster average*.

Definition 2.3.1. *Cluster Average*

For i.i.d. random vectors $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$, *the* i^{th} *cluster average is:*

$$W_i = \frac{1}{k}\sum_{j=1}^{k} X_{ij}.$$

Note that since the random vectors are i.i.d., it follows that W_1, W_2, ..., W_m are i.i.d.

The sample mean is still just the average of all the observed values.

Definition 2.3.2. *Cluster Sample Mean*

For i.i.d. random vectors $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$, *the cluster sample mean is:*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k} X_{ij}$$

Since clusters are all the same size, the cluster sample mean is equivalent to the sample mean of W.

Theorem 2.3.1. *The Cluster Sample Mean is the Sample Mean of Cluster Averages*

For i.i.d. random vectors $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$,

$$\overline{X} = \overline{W}.$$

Proof: Let $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$ be i.i.d. random vectors. Then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k} X_{ij} = \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k} X_{ij} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{k}\sum_{j=1}^{k} X_{ij} = \frac{1}{m}\sum_{i=1}^{m} W_i = \overline{W}. \ \square$$

The sample mean is now unbiased for what we'll call the *population mean cluster average*, i.e., the population mean of W, $\mathrm{E}[W]$.

Theorem 2.3.2. *The Cluster Sample Mean is Unbiased for the Population Mean Cluster Average*

For i.i.d. random vectors $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$,

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}[W].$$

Proof: Let $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$ be i.i.d. random vectors. Then

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}\left[\overline{W}\right] = \mathrm{E}[W],$$

where the second equality follows from Theorem 2.1.1. □

Characterizing the sampling variance becomes a bit more complicated, since we now need to account for the covariance between every pair of units within each cluster.

Theorem 2.3.3. *Sampling Variance of the Cluster Sample Mean*

For i.i.d. random vectors $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$, *the sampling variance of the cluster sample mean is:*

$$V\left(\overline{X}\right) = \frac{1}{mk^2}\sum_{j=1}^{k}\sum_{j'=1}^{k}\mathrm{Cov}\left(X_{ij}, X_{ij'}\right).$$

Proof: Let $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$ be i.i.d. random vectors. Then

$$V\left(\overline{X}\right) = V\left(\overline{W}\right) = \frac{V(W)}{m} = \frac{1}{m}V\left(\frac{1}{k}\sum_{j=1}^{k}X_{ij}\right) = \frac{1}{mk^2}V\left(\sum_{j=1}^{k}X_{ij}\right) = \frac{1}{mk^2}\sum_{j=1}^{k}\sum_{j'=1}^{k}\mathrm{Cov}\left(X_{ij}, X_{ij'}\right),$$

where the second equality follows from Theorem 2.1.2. □

A version of the WLLN still holds, but it now says that the cluster sample mean is consistent for the population mean cluster average. This result (and the following results) will require that the number of clusters $m \to \infty$, with k fixed. If the number of units per cluster $k \to \infty$, no such guarantees will hold.

Theorem 2.3.4. *WLLN for Cluster Samples*

Let $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$ *be i.i.d. random vectors with finite* $\mathrm{E}[X_{ij}]$ *and finite* $V(X_{ij}) > 0$, $\forall j$. *Then*

$$\overline{X} \xrightarrow{p} \mathrm{E}[W].$$

Proof: Let $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$ be i.i.d. random vectors with finite $\mathrm{E}[X_{ij}]$ and finite $V(X_{ij}) > 0$, $\forall j$. It is straightforward to show that this implies finite $\mathrm{E}[W]$ and finite $V(W)$.[67] Then by the WLLN, $\overline{X} = \overline{W} \xrightarrow{p} \mathrm{E}[W]$. □

[67] It could be the case that $V(W) = 0$, but then W is constant, $W = \mathrm{E}[W]$, and so it is trivially true that $\overline{X} = \overline{W} \xrightarrow{p} \mathrm{E}[W]$.

The CLT also still applies, though the standardized sample mean no longer reduces to $\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{\sigma}$. Instead, let $\mu_W = \mathrm{E}\left[W\right]$ and $\sigma_W = \sigma(W)$. Then

$$Z = \frac{\left(\overline{X} - \mathrm{E}\left[\overline{X}\right]\right)}{\sigma\left(\overline{X}\right)} = \frac{\left(\overline{W} - \mu_W\right)}{\sigma\left(\overline{W}\right)} = \frac{\left(\overline{W} - \mu_W\right)}{\sqrt{V\left(\overline{W}\right)}} = \frac{\left(\overline{W} - \mu_W\right)}{\sqrt{\frac{V(W)}{m}}} = \frac{\sqrt{m}\left(\overline{W} - \mu_W\right)}{\sigma_W} = Z_W,$$

or the standardized sample mean of W.

Theorem 2.3.5. *CLT for Cluster Samples*

Let $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$ *be i.i.d. random vectors such that* $\mathrm{E}\left[W\right]$ *and* $V(W) > 0$ *are finite, and let* Z *be the standardized sample mean. Then*

$$Z \xrightarrow{d} N(0,1),$$

or equivalently,

$$\sqrt{m}\left(\overline{X} - \mu_W\right) \xrightarrow{d} N(0, \sigma_W^2).$$

Proof: Let $(X_{11}, X_{12}, ...X_{1k})$, $(X_{21}, X_{22}, ...X_{2k})$, ...,$(X_{m1}, X_{m2}, ...X_{mk})$ be i.i.d. random vectors such that $\mathrm{E}\left[W\right]$ and $V(W) > 0$ are finite, and let Z be the standardized sample mean. Then by the CLT, $Z = Z_W \xrightarrow{d} N(0,1)$. And equivalently, $\sqrt{m}\left(\overline{X} - \mu_W\right) = \sqrt{m}\left(\overline{W} - \mu_W\right) \xrightarrow{d} N(0, \sigma_W^2)$. □

2.3.2 Inference with Clustering

The sampling variance of the cluster sample mean can be estimated using a plug-in estimator:

$$\hat{V}\left(\overline{X}\right) = \hat{V}\left(\overline{W}\right) = \frac{\hat{V}(W)}{m},$$

where $\hat{V}(W)$ is the unbiased sample variance of W. By Theorem 2.1.9 and the CMT, this estimator is consistent:

$$n\hat{V}\left(\overline{X}\right) = k\left[m\hat{V}\left(\overline{W}\right)\right] \xrightarrow{p} k\left[mV\left(\overline{W}\right)\right] = nV\left(\overline{X}\right).$$

Since $\overline{X}$ is asymptotically normal, normal approximation-based confidence intervals and p-values can then be obtained in the usual way.

How does clustering affect the efficiency of our estimates? Let's impose a little working assumption for illustrative purposes. Suppose that:

- $\forall i, j, V(X_{ij}) = \sigma^2$.
- $\forall i, j, j', \rho(X_{ij}, X_{ij'}) = \rho > 0$.

Then $\forall i, j, j', \text{Cov}\left(X_{ij}, X_{ij'}\right) = \rho(X_{ij}, X_{ij'})\sigma(X_{ij})\sigma(X_{ij'}) = \rho\sigma^2$, so

$$V\left(\overline{X}\right) = \frac{1}{mk^2}\sum_{j=1}^{k}\sum_{j'=1}^{k}\text{Cov}\left(X_{ij}, X_{ij'}\right) = \frac{1}{mk^2}\left(k\sigma^2 + k(k-1)\rho\sigma^2\right) = \frac{\sigma^2}{mk}\left(1 + \rho(k-1)\right).$$

Let's now consider two extreme cases. Suppose $\rho = 0$, i.e., no *intraclass correlation*. Then

$$V\left(\overline{X}\right) = \frac{\sigma^2}{mk} = \frac{\sigma^2}{n},$$

just as if we had no clustering. With no intraclass correlation, clustering induces no variance inflation, so it's as though we have n i.i.d. observations.

Now suppose $\rho = 1$, i.e., perfect intraclass correlation. Then

$$V\left(\overline{X}\right) = \frac{\sigma^2}{mk}\left(1 + k - 1\right) = \frac{\sigma^2}{mk}k = \frac{\sigma^2}{m}.$$

With perfect intraclass correlation, it's as though we only have m observations. Usually, the answer lies somewhere in between. We don't have to hypothesize this though, we can directly estimate the variance from the data.

What can we do if we want to conduct inference using some weird estimator and we don't have (or don't want to derive) an analytic estimator of the sampling variance of that estimator. Use the bootstrap! Specifically, with clustering, the appropriate method is the *block bootstrap*, which is exactly like the regular bootstrap except that now we resample whole *clusters* rather than individual units. Since we drew a sample of m i.i.d. clusters to begin with, we are again just replicating the sampling process, taking with-replacement samples of m clusters from our sample. The rest of the procedure is the same as before.

3 Regression

Truth is much too complicated to allow anything but approximations.

— JOHN VON NEUMANN

In this chapter, we discuss a simple and powerful tool for empirical practice: linear regression. Linear regression can be seen as a method for estimating the BLP using the data in your sample, much as the sample mean is an estimator for the population mean. But despite its simplicity, regression is a remarkably flexible tool that allows researchers to approximate the CEF in a principled and transparent manner.[68]

3.1 Regression as Plug-in Estimator

We begin by showing how linear regression is a simple plug-in estimator for the BLP. We know from Section 2.1.6 that plug-in estimators typically have good properties, and the linear regression estimator is no exception.

3.1.1 Bivariate Regression

Let's first consider the bivariate case. Recall that the BLP of Y given X is $g(X) = \alpha + \beta X$, where

$$\begin{aligned}\alpha &= \mathrm{E}\left[Y\right] - \frac{\mathrm{Cov}\left(X, Y\right)}{V(X)}\mathrm{E}\left[X\right], \\ \beta &= \frac{\mathrm{Cov}\left(X, Y\right)}{V(X)}.\end{aligned}$$

Applying Theorems 1.4.5 and 1.4.12, these equations can be rewritten in terms of expected values:

$$\begin{aligned}\alpha &= \mathrm{E}\left[Y\right] - \frac{\mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2}\mathrm{E}\left[X\right] \\ \beta &= \frac{\mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]}{\mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2}\end{aligned}$$

We can then *estimate* the BLP using simple plug-in estimation:

$$\begin{aligned}\hat{\alpha} &= \overline{Y} - \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}\overline{X}, \\ \hat{\beta} &= \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}.\end{aligned}$$

[68] For a more technical treatment of the results in this chapter, we recommend Hansen (2013).

Thus, we can use *sample* data to estimate the population relationship: $\hat{\alpha}$ is the regression estimate of the intercept of the BLP and is consistent for α, and $\hat{\beta}$ is the regression estimate of the slope of the BLP and is consistent for β. (Consistency follows immediately from the WLLN and CMT.)

Example 3.1.1. *Estimating the BLP*

Suppose we have a random vector (X, Y) such that

$$Y = 1 + 2X + 5\epsilon,$$

where X and ϵ are independent and each distributed as $N(0, 1)$. This fully describes the joint distribution of (X, Y), so we can derive the CEF of Y given X:

$$\mathrm{E}[Y|X] = \mathrm{E}[1 + 2X + 5\epsilon|X] = 1 + 2\mathrm{E}[X|X] + 5\mathrm{E}[\epsilon] = 1 + 2X.$$

In this example, the CEF is linear, so the BLP *is* the CEF, and so we have $\alpha = 1$, $\beta = 2$. Now, suppose we didn't know the joint distribution of (X, Y) and we wanted to estimate the CEF of Y given X from sample data. We would observe a sample of draws from (X, Y); such a sample is shown in Table 3.1.1.

Unit	y	x
1	6.88	1.37
2	1.17	0.30
3	16.16	1.57
4	-4.84	-0.69
5	-6.17	-0.64
6	0.67	0.08
7	3.36	0.47
8	-0.69	1.33
9	8.03	-0.13
10	-3.35	0.59
11	1.87	-0.44
12	13.54	0.76
13	4.06	0.73
14	6.72	-3.54
15	7.20	-0.05

Table 3.1.1 *Sample of* $n = 15$ *Draws from* (X, Y)

That's all we have. How would we try to estimate the CEF from this data? Well, the CEF is hard to characterize; it could have *any* functional form. (In this case, we know that it is in fact linear, but we're pretending we don't know that.) Let's use a linear approximation. Regression consistently estimates the MMSE *linear* approximation to the CEF, i.e., the BLP.

In Figure 3.1.1, we overlay the data points, the actual (unknown to the researcher) CEF and the estimate of the BLP. We see that, even with an n of 15, we approximate the relationship fairly well. Naturally, if we were to add more data, this approximation would improve accordingly.

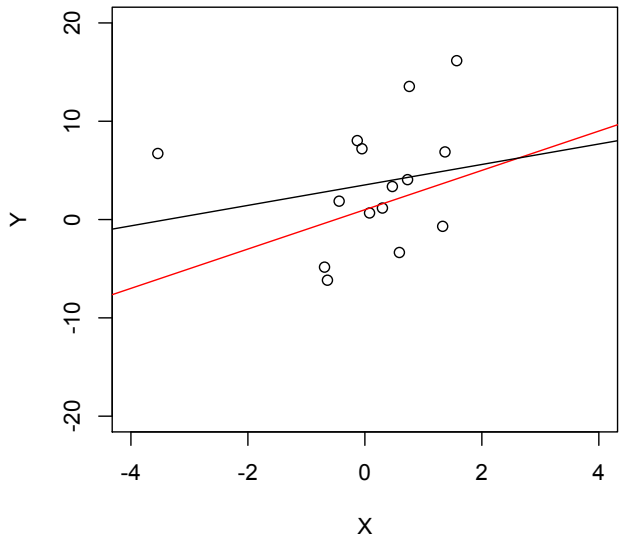


Figure 3.1.1. Illustration of $n = 15$ Sample of Draws from (X, Y). Overlaid over the data points is the CEF in red, and the regression estimate in black.

3.1.2 Multivariate Regression

Suppose now that we have K explanatory variables: X_1, X_2, ..., X_K.[69] In this case, as we saw in Section 1.4.9, the BLP is still defined as the linear function that minimizes MSE. I.e., the BLP (also known as the *population regression function*) is the linear function $g(X_1, X_2, ..., X_K) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K$ such that

$$(\beta_0, \beta_1, ..., \beta_K) = \underset{(b_0, b_1, ..., b_K) \in \mathbb{R}^{K+1}}{\arg\min} \mathrm{E}\left[\left[Y - (b_0 + b_1 X_1 + b_2 X_2 + ... b_K X_K)\right]^2\right].$$

How would we estimate the BLP? Unfortunately, in the multivariate case we no longer have the nice formulas $\alpha = \mathrm{E}[Y] - \frac{\mathrm{Cov}(X,Y)}{V(X)}\mathrm{E}[X]$ and $\beta = \frac{\mathrm{Cov}(X,Y)}{V(X)}$. But the plug-in principle still applies; we can still replace expectations with sample means in the definition of the BLP to obtain an estimate thereof.

Definition 3.1.1. *Linear Regression Estimator*

Given n i.i.d. draws of a random vector $(X_1, X_2, ..., X_K, Y)$, where each observation i is denoted by $(X_{1i}, X_{2i}, ..., X_{Ki}, Y_i)$, the linear regression estimator is the function $\hat{g}(X_1, X_2, ..., X_K) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_K X_K$ such that

$$(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K) = \underset{(b_0, b_1, ..., b_K) \in \mathbb{R}^{K+1}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i} + ... b_K X_{Ki})\right]^2.$$

[69] We use the term "multivariate regression" to denote regression with a random variable Y as the outcome but with more than one explanatory variable. Others use it to note regression models where the outcome is a random vector.

The quantity $Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ...\hat{\beta}_K X_{Ki})$ is known as the *i^{th} residual*. It is the difference between the observed value of Y_i and the value we would have predicted given the observed values of X_{1i}, X_{2i}, ..., X_{Ki} and our estimate of the BLP. A residual is the sample analogue of a prediction *error*, which is the difference between an observed value of Y_i and the value that would have been predicted based on the observed values of X_{1i}, X_{2i}, ..., X_{Ki} and some actual population predictor (e.g, the true CEF or BLP). Linear regression is also known as *ordinary least squares* (OLS), since it finds the fit that the minimizes the mean of squared residuals (or, equivalently, minimizes the sum of squared residuals).

Intuitively, since the sample mean converges to the expected value, the solution to the problem with sample data converges to the solution to the problem for the population. This general principle usually* works for most statistical problems: write down the population problem, and then solve it as though the sample data were the population distribution.

Given suitable regularity conditions, the WLLN and the CMT imply that

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K) &= \underset{(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_K)\in\mathbb{R}^{K+1}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ...\hat{\beta}_K X_{Ki})\right]^2 \\ &\xrightarrow{p} \underset{(\beta_0,\beta_1,\beta_2,...,\beta_K)\in\mathbb{R}^{K+1}}{\arg\min} \mathrm{E}\left[[Y - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_K X_K)]^2\right] \\ &= (\beta_0, \beta_1, ..., \beta_K).\end{aligned}$$

I.e., the linear regression estimator is consistent for the BLP.[70] In the following sections, we show two different methods for computing the regression estimator.

3.1.3 Regression with Matrix Algebra

Suppose we have n i.i.d. draws of some random vector $(X_1, X_2, ..., X_K, Y)$, with each observation i denoted by $(X_{1i}, X_{2i}, ..., X_{Ki}, Y_i)$. For any $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K) \in \mathbb{R}^{K+1}$ and for every observation i, we can define the i^{th} residual as

$$e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_K X_{Ki}),$$

so that

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_K X_{Ki} + e_i.$$

Define[71]

$$\mathbb{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{Kn} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \text{ and } \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{pmatrix}.$$

[70] To be very formal, we further note that, given any reasonable definition of distance between functions, $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K) \xrightarrow{p} (\beta_0, \beta_1, ..., \beta_K)$ implies $\hat{g}(X_1, X_2, ..., X_K) \xrightarrow{p} g(X_1, X_2, ..., X_K)$.

[71] All matrices in this book shall be denoted by "blackboard bold" letters, e.g., $\mathbb{X}$, $\mathbb{W}$, etc.

Then the system of equations

$$\begin{aligned} Y_1 &= \hat{\beta}_0 + \hat{\beta}_1 X_{11} + \hat{\beta}_2 X_{21} + ...\hat{\beta}_K X_{K1} + e_1 \\ Y_2 &= \hat{\beta}_0 + \hat{\beta}_1 X_{12} + \hat{\beta}_2 X_{22} + ...\hat{\beta}_K X_{K2} + e_2 \\ &\vdots \\ Y_n &= \hat{\beta}_0 + \hat{\beta}_1 X_{1n} + \hat{\beta}_2 X_{2n} + ...\hat{\beta}_K X_{Kn} + e_n \end{aligned}$$

can be rewritten as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{Kn} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix},$$

or equivalently,

$$\mathbf{Y} = \mathbb{X}\hat{\boldsymbol{\beta}} + \mathbf{e}.$$

To compute the linear regression estimator, we want to find

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{K+1}}{\arg\min} \sum_{i=1}^{n} e_i^2 = \underset{\boldsymbol{\beta} \in \mathbb{R}^{K+1}}{\arg\min} \left(\mathbf{e}^T \mathbf{e} \right) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{K+1}}{\arg\min} (\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}})^T (\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}).$$

The first order condition is

$$-2\mathbb{X}^T(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}) = \mathbf{0},$$ [72]

which is equivalent to

$$\mathbb{X}^T \mathbf{e} = \mathbf{0}.$$

So we have

$$\begin{aligned} \mathbf{Y} &= \mathbb{X}\hat{\boldsymbol{\beta}} + \mathbf{e} \\ \mathbb{X}^T\mathbf{Y} &= \mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\beta}} + \mathbb{X}^T\mathbf{e} \\ \mathbb{X}^T\mathbf{Y} &= \mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\beta}} + \mathbf{0} \\ \mathbb{X}^T\mathbf{Y} &= \mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\beta}} \\ \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbf{Y} &= \hat{\boldsymbol{\beta}}. \end{aligned}$$

[72] If you don't know any matrix calculus, don't worry. Just take our word for it that the derivative of $(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}})$ with respect to $\hat{\boldsymbol{\beta}}$ is $-2\mathbb{X}^T(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}})$, noticing how this resembles the chain rule. Note that $\mathbf{0}$ denotes the zero vector of appropriate length.

Thus, $\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbf{Y}$ is the coefficient vector of the linear regression estimator. Two noteworthy properties of the regression estimator also follow from this solution. The first order condition $\mathbb{X}^T\mathbf{e} = \mathbf{0}$ is equivalent to

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{K1} & X_{K2} & \cdots & X_{Kn} \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

which summarizes the system of equations

$$\begin{aligned} e_1 + e_2 + ... + e_n &= 0 \\ e_1X_{11} + e_2X_{12} + ... + e_nX_{1n} &= 0 \\ e_1X_{21} + e_2X_{22} + ... + e_nX_{2n} &= 0 \\ &\vdots \\ e_1X_{K1} + e_2X_{K2} + ... + e_nX_{Kn} &= 0. \end{aligned}$$

The first equation says that the sum of the residuals is zero. Intuitively, if the errors did not sum to zero, we could obtain a better fit by changing the constant term. Furthermore, dividing each equation by n,

$$\begin{aligned} \overline{e} &= \frac{1}{n}\sum_{i=1}^{n} e_i = 0 \\ \overline{eX_1} &= \frac{1}{n}\sum_{i=1}^{n} e_iX_{1i} = 0 \\ \overline{eX_2} &= \frac{1}{n}\sum_{i=1}^{n} e_iX_{2i} = 0 \\ &\vdots \\ \overline{eX_K} &= \frac{1}{n}\sum_{i=1}^{n} e_iX_{Ki} = 0, \end{aligned}$$

which implies that for $k = 1, 2, ..., K$, the sample covariance between the residuals and X_k is zero:

$$\overline{eX_k} - \overline{e}\,\overline{X}_k = 0.$$

3.1.4 Regression Using the Frisch-Waugh-Lovell Theorem

The OLS coefficient estimates for multivariate regression can also be obtained through iterated application of the bivariate regression formulas from Section 3.1.1, using a procedure known as *partial regression*.

This result is a direct consequence of the *Frisch-Waugh-Lovell Theorem* (FWL Theorem), which we do not formally state here, though a straightforward statement and proof are available in Lovell (2008).

Suppose we have n i.i.d. draws of a random vector (X_1, X_2, Y), with each observation i denoted by (X_{1i}, X_{2i}, Y_i). Suppose we perform the following procedure:

- Regress Y on X_2, i.e., compute the linear regression estimator $\hat{g}_Y(X_2) = \hat{\alpha}_0 + \hat{\alpha}_2 X_2$, using the bivariate plug-in estimators.
- Regress X_1 on X_2, i.e., compute the linear regression estimator $\hat{g}_{X_1}(X_2) = \hat{\gamma}_0 + \hat{\gamma}_2 X_2$, using the bivariate plug-in estimators.
- Compute the residuals from these two regressions: $\forall i \in \{1, 2, ..., n\}$,

$$Y_i^r = Y_i - \hat{g}_Y(X_{2i}),$$
$$X_{1i}^r = X_{1i} - \hat{g}_{X_1}(X_{2i}),$$

- Regress Y^r on X_1^r, i.e., compute the linear regression estimator $\hat{g}_{Y^r}(X_1^r) = \hat{\beta}_0 + \hat{\beta}_1 X_1^r$, using the bivariate plug-in estimators. Or, equivalently (since we have already residualized off the constant), regress Y^r on X_1^r *without a constant*, i.e., compute the linear regression estimator $\hat{g}_{Y^r}(X_1^r) = \hat{\beta}_1 X_1^r$. In this case, the bivariate plug-in estimator[73] is simply

$$\hat{\beta}_1 = \frac{\overline{Y^r X_1^r}}{\overline{X_1^{r2}}}.$$

Then the FWL Theorem says that the $\hat{\beta}_1$ resulting from this last regression is equal to the $\hat{\beta}_1$ for the multivariate regression estimator $\hat{g}_Y(X_1, X_2) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$. Note that if we instead first regress Y and X_2 on X_1, and then regress the residuals Y^r on X_2^r (again, with or without a constant), the coefficient on X_2^r in the final regression will be the $\hat{\beta}_2$ for the multivariate regression estimator.

This procedure can even be applied recursively when you have more than two regressors. For example, with three regressors, X_1, X_2, and X_3, you can first "partial out" X_3 by computing the bivariate regressions of Y, X_1, and X_2 on X_3, then regress the resulting residuals Y^r and X_1^r on X_2^r, and finally perform a bivariate regression with the residuals Y^{rr} and X_1^{rr} from that regression. The coefficient on X_1^{rr} in the final regression will be the $\hat{\beta}_1$ for the multivariate regression of Y on X_1, X_2, and X_3. Again, changing the order in which we partial out the Xs, we can likewise obtain $\hat{\beta}_2$ and $\hat{\beta}_3$ for the multivariate regression. The same logic applies for any number of regressors.

We now present a simplified proof sketch of the FWL Theorem.[74] Suppose we have n i.i.d. draws of a random vector $(X_1, X_2, ..., X_K, Y)$, with each observation i denoted by $(X_{1i}, X_{2i}, ..., X_{Ki}, Y_i)$. Let $\mathbb{W}$

[73] This can be derived by solving the first order condition to minimize MSE, as in Chapter 1, and then plugging in sample means for expected values. We leave this as an exercise to the reader.

[74] Adapted from Davidson and MacKinnon (2004), Section 2.4.

refer to the first K columns of $\mathbb{X}$, and let $\mathbf{X}_K$ refer to the $(K+1)^{\text{th}}$, i.e.

$$\mathbb{W} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{(K-1)1} \\ 1 & X_{12} & X_{22} & \cdots & X_{(K-1)2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{(K-1)n} \end{pmatrix}, \quad \mathbf{X}_K = \begin{pmatrix} X_{K1} \\ X_{K2} \\ \vdots \\ X_{Kn} \end{pmatrix}.$$

We're now going to partial out X_1,X_2, ..., X_{K-1} and then show that partial regression recovers the multiple regression coefficient on X_K.

- Regressing Y on X_1,X_2, ..., X_{K-1}, we obtain the coefficient vector $\hat{\boldsymbol{\alpha}} = (\mathbb{W}^T\mathbb{W})^{-1}\mathbb{W}^T\mathbf{Y}$.
- Regressing X_K on X_1,X_2, ..., X_{K-1}, we obtain the coefficient vector $\hat{\boldsymbol{\gamma}} = (\mathbb{W}^T\mathbb{W})^{-1}\mathbb{W}^T\mathbf{X}_K$.
- The residual vectors from these two regressions are then

$$\begin{aligned} \mathbf{Y}^r &= \mathbf{Y} - \mathbb{W}\hat{\boldsymbol{\alpha}}, \\ \mathbf{X}_K^r &= \mathbf{X}_K - \mathbb{W}\hat{\boldsymbol{\gamma}}. \end{aligned}$$

- Regressing Y^r on X_K^r without a constant, we obtain $\hat{\beta}_K = \frac{\overline{Y^r X_K^r}}{\overline{X_K^r}^2}$, with associated vector of residuals

$$\mathbf{e} = \mathbf{Y}^r - \mathbf{X}_K^r\hat{\beta}_K.$$

Applying the first order condition derived in Section 3.1.3,

- $\mathbb{W}^T\mathbf{e} = \mathbb{W}^T\left(\mathbf{Y}^r - \mathbf{X}_K^r\hat{\beta}_K\right) = \mathbb{W}^T\mathbf{Y}^r - \mathbb{W}^T\mathbf{X}_K^r\hat{\beta}_K = \mathbf{0} + \mathbf{0}\hat{\beta}_K = \mathbf{0}$.
- $\mathbf{X}_K^T\mathbf{e} = \left(\mathbf{X}_K^r + \mathbb{W}\hat{\boldsymbol{\gamma}}\right)^T\mathbf{e} = {\mathbf{X}_K^r}^T\mathbf{e} + \hat{\boldsymbol{\gamma}}^T\mathbb{W}^T\mathbf{e} = 0 + \hat{\boldsymbol{\gamma}}^T\mathbf{0} = 0$.

So we have

$$\begin{aligned} e_1 + e_2 + ... + e_n &= 0 \\ e_1X_{11} + e_2X_{12} + ... + e_nX_{1n} &= 0 \\ e_1X_{21} + e_2X_{22} + ... + e_nX_{2n} &= 0 \\ &\vdots \\ e_1X_{K1} + e_2X_{K2} + ... + e_nX_{Kn} &= 0, \end{aligned}$$

which is equivalent to $\mathbb{X}^T\mathbf{e} = \mathbf{0}$, the first order condition for the full multiple regression.

Thus,

$$\begin{aligned} \mathbf{Y} &= \mathbb{W}\hat{\boldsymbol{\alpha}} + \mathbf{Y}^r \\ &= \mathbb{W}\hat{\boldsymbol{\alpha}} + \mathbf{X}_K^r\hat{\beta}_K + \mathbf{e} \\ &= \mathbb{W}\hat{\boldsymbol{\alpha}} + (\mathbf{X}_K - \mathbb{W}\hat{\boldsymbol{\gamma}})\hat{\beta}_K + \mathbf{e} \\ &= \mathbb{W}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\gamma}}\hat{\beta}_K) + \mathbf{X}_K\hat{\beta}_K + \mathbf{e}, \text{ with } \mathbb{X}^T\mathbf{e} = \mathbf{0}, \end{aligned}$$

so β_K is the coefficient on X_K for the full multiple regression. $\square$

3.2 Inference

We have shown that we can estimate the BLP using simple plug-in estimation. Just like every other estimator we've seen, each of our coefficient estimators $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_K$ is a random variable. Each therefore has a sampling distribution and, thus, a sampling variance. As usual, we want to *estimate* these sampling variances so that we can construct confidence intervals and perform hypothesis tests for the values of these coefficients.

3.2.1 Standard Errors

Let's first consider the simplest case. Suppose that we had no explanatory variables. I.e., suppose we wanted to estimate

$$\beta_0 = \underset{b_0 \in \mathbb{R}}{\arg\min}\, \mathrm{E}\left\{[Y - b_0]^2\right\}.$$

We know from Theorem 1.4.11 that the solution is

$$\beta_0 = \mathrm{E}[Y].$$

The sample analogue to β_0 would then be $\overline{Y}$, so using plug-in estimation,

$$\hat{\beta}_0 = \overline{Y}.$$

Thus, the linear regression estimate of β_0 with no explanatory variables is the sample mean. So inference proceeds exactly as we saw in Chapter 2: the standard error of $\hat{\beta}_0$ is

$$\sigma\left(\hat{\beta}_0\right) = \sqrt{V\left(\hat{\beta}_0\right)} = \sqrt{V\left(\overline{Y}\right)} = \sqrt{\frac{V(Y)}{n}},$$

which we can estimate with

$$\hat{\sigma}\left(\hat{\beta}_0\right) = \sqrt{\hat{V}\left(\hat{\beta}_0\right)} = \sqrt{\hat{V}\left(\overline{Y}\right)} = \sqrt{\frac{\hat{V}(Y)}{n}}.$$

As before, we could use the estimated sampling variance to compute a 95% confidence interval for the population mean, β_0:

$$\widehat{CI}_{.95}(\beta_0) = \left(\hat{\beta}_0 - 1.96\,\hat{\sigma}\left(\hat{\beta}_0\right), \hat{\beta}_0 + 1.96\,\hat{\sigma}\left(\hat{\beta}_0\right)\right).$$

Of course, estimating standard errors becomes more complicated when we have one or more regressors, but the general *principles* for estimating uncertainty, confidence intervals, and p-values translate from the sample mean to the regression estimator. For each coefficient β_k, it is possible to derive an estimate of the standard error of $\hat{\beta}_k$:

$$\hat{\sigma}\left(\hat{\beta}_k\right) = \sqrt{\hat{V}(\hat{\beta}_k)}.$$

You should (almost always) use *robust standard errors*. These provide consistent estimates of the true standard errors under i.i.d. sampling. We show how to derive robust standard errors in the following

sections, but "canned" commands can do this very straightforwardly in most statistical software. Again, we can then use these standard error estimates to compute normal approximation based confidence intervals and p-values, as before.[75] E.g., suppose we had a random vector (X_1, X_2, Y) with the BLP

$$g(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

and suppose we obtained the following *estimates* after fitting the linear regression:

$$\begin{aligned}
\hat{\beta}_0 = 0.60, \hat{\sigma}\left(\hat{\beta}_0\right) = 0.25, \\
\hat{\beta}_1 = 0.75, \hat{\sigma}\left(\hat{\beta}_1\right) = 1.25, \\
\hat{\beta}_2 = 3.00, \hat{\sigma}\left(\hat{\beta}_2\right) = 1.00.
\end{aligned}$$

We could then construct the following 95% confidence intervals under normal approximation:

$$\begin{aligned}
\widehat{CI}_{.95}(\beta_0) = (0.11, 1.09), \\
\widehat{CI}_{.95}(\beta_1) = (-1.70, 3.20), \\
\widehat{CI}_{.95}(\beta_2) = (1.04, 4.96).
\end{aligned}$$

Thus, for instance, we can say approximately with 95% confidence (if n is large) that the *conditional* slope of the BLP with respect to X_2 lies in the interval (1.04, 4.96).[76]

Likewise, normal approximation based p-values can be computed in the manner shown in Chapter 2. So for example, we can state that, under the null hypothesis that β_2 is zero, the probability that we would have observed a $\hat{\beta}_2$ as extreme as we did is less than 0.05.

3.2.2 Robust Standard Errors with Matrix Algebra

We now present the derivation of the robust standard error estimator for $\hat{\boldsymbol{\beta}}$. Define the vector of errors as the differences between the observed values of Y and the (true) BLP of Y given the observed values of X_1, X_2, ..., X_K:

$$\boldsymbol{\epsilon} = \mathbf{Y} - \mathbb{X}\boldsymbol{\beta}.$$

We can then decompose $\hat{\boldsymbol{\beta}}$ as

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbf{Y} \\
&= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T(\mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{X}\boldsymbol{\beta} + \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\boldsymbol{\epsilon}.
\end{aligned}$$

[75] It can be shown that, under suitable regularity conditions, the regression coefficient estimators are asymptotically normal.

[76] I.e., if we repeatedly sampled n observations from (X_1, X_2, Y), our interval would cover the true *conditional* slope of the BLP 95% percent of the time.

Thus,

$$\begin{aligned} V\left(\hat{\boldsymbol{\beta}}\right) &= V\left[\boldsymbol{\beta} + \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\boldsymbol{\epsilon}\right] \\ &= V\left[\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\boldsymbol{\epsilon}\right]. \end{aligned}$$

By Slutsky's Theorem,

$$\begin{aligned} nV\left(\hat{\boldsymbol{\beta}}\right) &\approx nV\left[\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1}\mathbb{X}^T\boldsymbol{\epsilon}\right] \\ &= n\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1} V\left(\mathbb{X}^T\boldsymbol{\epsilon}\right)\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1} \\ &= n\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1}\mathrm{E}\left[\mathbb{X}^T\mathrm{diag}(\boldsymbol{\epsilon}^2)\mathbb{X}\right]\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1}, \end{aligned}$$

where the quality of the approximation improves as $n \to \infty$.

The idea is that we can characterize the sampling variability of the estimator just by characterizing the relationship between the errors and the explanatory variables. To estimate $V(\hat{\boldsymbol{\beta}})$, we simply plug in residuals for errors and observed values for expected values. The resulting plug-in estimator is known as the *sandwich estimator*:

$$\hat{V}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathrm{diag}\left[(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}})^2\right]\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}.$$

The sandwich estimator uses the residuals to *estimate* the relationship between the X variables and the errors. The robust standard error estimator is then given by

$$\hat{\sigma}\left(\hat{\boldsymbol{\beta}}\right) = \sqrt{\hat{V}\left(\hat{\boldsymbol{\beta}}\right)}.$$

Note that this is an *asymptotic* approximation of the standard error. Without assumptions that we would not be willing to believe (e.g., normality of errors, fixed regressors), we cannot derive an exact, closed-form expression for the standard error of our regression estimates given any finite n.

3.2.3 Robust Standard Errors using the Frisch-Waugh-Lovell Theorem

Once again, we can obtain some insights regarding the robust standard error estimator by deriving it using the FWL Theorem. Consider the bivariate plug-in estimator for partial regression *without a constant*:

$$\hat{\beta}_K = \frac{\overline{Y^r X_K^r}}{\overline{X_K^r}^2} = \frac{\sum_{i=1}^n Y_i^r X_{Ki}^r}{\sum_{i=1}^n (X_{Ki}^r)^2} = \frac{\sum_{i=1}^n e_i X_{Ki}^r}{\sum_{i=1}^n (X_{Ki}^r)^2} + \beta_K,$$

where $\forall i \in \{1, 2, ..., n\}, Y_i^r = \beta_K X_{Ki}^r + e_i$.

Then, since $\hat{\mathrm{E}}\left[(\hat{X}_i^r)^2\right]$ (and all of the $\hat{\beta}$s that generate the residuals, where appropriate) converge "fairly quickly":

$$V[\hat{\beta}] = V\left[\frac{\sum_{i=1}^n e_i X_i^r}{\sum_{i=1}^n (X_i^r)^2}\right] \approx \frac{1}{\mathrm{E}\left[(X_i^r)^2\right]^2}\frac{V[e_i X_i^r]}{n} = \frac{1}{n}\frac{V[e_i X_i^r]}{V[X_i^r]^2},$$

where this is a good approximation with large n.[77]

To estimate $V[\hat{\beta}]$, we will use our standard strategy of using a plug-in estimator:

$$\hat{V}[\hat{\beta}] = \frac{1}{n}\frac{\hat{V}[e_i X_i^r]}{\hat{V}[X_i^r]^2},$$

where the residual $e_i = Y_i^r - X_i^r\hat{\beta}$. This estimator is logically equivalent to the standard (HC0) Huber-White "robust" variance estimate. The square root of $\hat{V}[\hat{\beta}]$ is the robust standard error estimate.[78]

3.2.4 A Note on Collinearity and Micronumerosity

When two or more explanatory variables in a regression are highly correlated, they are said to be *collinear* (sometimes *multicollinear*). Econometricians and statisticians often worry about collinearity as it may increase the standard error associated with a given regression coefficient. We will not be so concerned with collinearity as a problem per se, but believe it is worth addressing as a feature of a given regression strategy.

First, we will show why collinearity often increases the standard error of a regression coefficient. The derivation using the FWL theorem in Section 3.2.3 provides some insight. Note that if we are concerned with the asymptotic efficiency in estimating $\hat{\beta}$, then $V[\hat{\beta}] \approx \frac{1}{n}\frac{V[e_i X_i^r]}{V[X_i^r]^2}$ tells us something about how collinearity affects standard errors (given that the standard error of $\hat{\beta}$ is just the square root of $V[\hat{\beta}]$). All else equal, as a general heuristic, we'd want to keep the variance of X_i^r *high* and the variance of e_i *low* When the variance of X_i^r is low, that implies that there is high collinearity—the explanatory variable of interest is very well-explained by the other explanatory variables. (I.e., the variance of the residuals from a regression of X_i on the other variables would be low if X_i were collinear with the other variables.)

This variance expression suggests a basic intuition for efficiency—we'd like explain the outcome well, but not at the cost of predicting the factor of interest too well. But changing the set of explanatory variables changes the inferential target. Adding a variable to the set of explanatory variables changes the coefficient from reflecting a marginal relationship to a conditional relationship. If we are interested in characterizing how X_i predicts Y_i holding X_i fixed, then including X_i in the regression is a necessity.

Collinearity is often a feature of the process under study. If there's not much variation in X_i once we have conditioned on the other factors, then we will incur a penalty in the size of the standard error. But there is one sure-fire way to shrink the standard error—collect more data. The problems introduced by collinearity are, in fact, logically equivalent to the problem of just having too few observations.

---

[77] A proof of this claim follows directly from Slutsky's Theorem. Since our interest is in $nV[\hat{\boldsymbol{\beta}}]$, we can equivalently write $nV[\hat{\beta}]$ as $V[\sqrt{n}\hat{\beta}] = V\left[\frac{\sqrt{n}\sum_{i=1}^n e_i X_i^r}{\sum_{i=1}^n (X_i^r)^2}\right]$. It then follows that the limiting distribution of $\sqrt{n}\hat{\beta}$ is governed by the numerator, and that the denominator can be treated as fixed.

[78] There are many other variants of the robust variance estimator, obtained by premultiplying the basic plug-in estimator (HC0) by $\frac{n}{n-K}$, $\frac{n}{n-K-2}$, etc. These are analogous to the unbiased sample variance as opposed to the plug-in sample variance. In the case of robust variance estimation, we have not just two but many estimators (HC0, HC1, HC2, etc.), but as with the unbiased vs. plug-in sample variance estimators, they all have the same probability limit.

Goldberger (1991, Ch. 23.3) draws this connection elegantly: "Econometrics texts devote many pages to the problem of multicollinearity in multiple regression, but they say little about the closely analogous problem of small sample size in estimation of a univariate mean. Perhaps that imbalance is attributable to the lack of an exotic polysyllabic name for 'small sample size'. If so, we can remove that impediment by introducing the term *micronumerosity*." Goldberger proceeds to (somewhat sarcastically) discuss the problems that micronumerosity poses in a manner analogous to textbook discussions of collinearity. But most importantly in our view, Goldberger notes (albeit in the context of a satire of a textbook discussion of collinearity), "If micronumerosity proves serious in the sense that the estimate of [a mean] has an unsatisfactorily low degree of precision, we are in the statistical position of not being able to make bricks without straw. The remedy lies essentially in the acquisition, if possible, of larger samples from the same population."

Suffice it to say that we agree with Goldberger. Sometimes uncertainty is fundamentally large, and a larger sample is the only sure-fire solution.

3.2.5 Classical Variance Estimation

There is an alternative to the robust variance estimator, which we will refer to as the *classical variance estimator*. Suppose we *assume* that $V(\epsilon|X_1, X_2, ..., X_K) = \sigma^2$ for all possible values of $X_1, X_2, ..., X_K$. This assumption is known as (conditional) *homoskedasticity*. Then

$$\begin{aligned} nV\left(\hat{\boldsymbol{\beta}}\right) &\approx n\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1}\mathrm{E}\left[\mathbb{X}^T\mathrm{diag}(\boldsymbol{\epsilon}^2)\mathbb{X}\right]\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1} \\ &= n\sigma^2\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1}\mathrm{E}\left[\mathbb{X}^T\mathbb{I}\mathbb{X}\right]\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1} \\ &= n\sigma^2\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1}\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1} \\ &= n\sigma^2\left(\mathrm{E}\left[\mathbb{X}^T\mathbb{X}\right]\right)^{-1}, \end{aligned}$$

where the approximation follows from Slutsky's Theorem, and the first equality follows from the fact that $\sigma^2 = V(\epsilon) = \mathrm{E}\left[\epsilon^2\right]$. The plug-in estimator for the variance is then

$$\hat{V}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2\left[\mathbb{X}^T\mathbb{X}\right]^{-1},$$

where $\hat{\sigma}^2 = \hat{V}(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}})$.

This is the variance estimator for regression given in most textbooks. Yet it relies on a strong assumption: that the magnitude of the errors (i.e., how far off the BLP is, on average, from the true value) is *unrelated* to the values of the Xs. This assumption is unlikely to hold; usually, in real life, the errors will be *heteroskedastic*, because real-life phenomena are messy. If homoskedasticity *does* hold, then in large samples, the classical and robust variance estimates will agree. With large n and i.i.d. sampling, always use robust standard errors to characterize the uncertainty of your regression fit.

3.2.6 The Bootstrap

As with all estimators, we can also estimate the sampling variability of the regression estimator straightforwardly using the bootstrap, as described in Section 2.2.3. E.g., to estimate $V(\hat{\beta}_1)$,

1. Take a *with replacement* sample of size n from your sample.
2. Calculate $\hat{\beta}_1$ using this bootstrap sample.
3. Repeat steps 1 and 2 many, many times.
4. Then $\hat{V}\left(\hat{\beta}_1\right)$ is just the sample variance of the resulting collection of bootstrap estimates.

As $n \to \infty$, robust standard errors and bootstrapped standard errors become very similar.

3.3 Clustering

How does clustering affect our inferences with regression? The answer is: in much the same way as it does for sample means. Suppose that we observe m i.i.d. clusters, each consisting of k observations, so that $n = mk$. So for each cluster i we have a random vector $(\mathbf{X}_{i1}, Y_{i1}, \mathbf{X}_{i2}, Y_{i2}, ..., \mathbf{X}_{ik}, Y_{ik})$. Thus, while observations must be independent across clusters—e.g., $(\mathbf{X}_{ij}, Y_{ij}) \perp\!\!\!\perp (\mathbf{X}_{i'j}, Y_{i'j})$—within clusters they may be statistically dependent—i.e., it is not generally the case that $(\mathbf{X}_{ij}, Y_{ij}) \perp\!\!\!\perp (\mathbf{X}_{ij'}, Y_{ij'})$.

Consider a "long" dataset that looked as follows:

$Cluster_i$	$Unit_j$	Y_{ij}	$X_{1ij}ij$	X_{2ij}
1	1	3	0	1
1	2	2	1	5
1	3	0	0	3
2	1	2	0	4
2	2	1	0	5
2	3	2	0	1
3	1	4	1	0
3	2	1	0	1
3	3	0	1	1
...	...	...	...	...

Then the question is, what adjustments do we need to make if we were to analyze the "long" dataset with standard regression methods?

Let's see how the variance now looks for our partialled out regression:

$$
\begin{aligned}
V[\hat{\beta}] &= V\left[\frac{\sum_{i=1}^{m}\sum_{j=1}^{k} e_{ij}X_{1ij}^{r}}{\sum_{i=1}^{m}\sum_{j=1}^{k}(X_{1ij}^{r})^2}\right] \\
&\approx \frac{1}{\mathrm{E}\left[(X_{1ij}^{r})^2\right]^2} V(\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k} e_{ij}X_{1ij}^{r}) \\
&= \frac{1}{\mathrm{E}\left[(X_{1ij}^{r})^2\right]^2} \frac{V(\sum_{i=1}^{m}\sum_{j=1}^{k} e_{ij}X_{1ij}^{r})}{n^2} \\
&= \frac{1}{\mathrm{E}\left[(X_{1ij}^{r})^2\right]^2} \frac{mV(\sum_{j=1}^{k} e_{ij}X_{1ij}^{r})}{n^2} \\
&= \frac{1}{\mathrm{E}\left[(X_{1ij}^{r})^2\right]^2} \frac{m\sum_{j=1}^{k}\sum_{j'=1}^{k}\mathrm{Cov}\left(e_{ij}X_{1ij}^{r}, e_{ij'}D_{ij'}^{r}\right)}{n^2} \\
&= \frac{1}{\mathrm{E}\left[(X_{1ij}^{r})^2\right]^2} \frac{mkV[e_{ij}X_{1ij}^{r}] + m\sum_{j=1}^{k}\sum_{j'\neq j}\mathrm{Cov}\left(e_{ij}X_{1ij}^{r}, e_{ij'}D_{ij'}^{r}\right)}{n^2}
\end{aligned}
$$

where $\mathrm{E}\left[\cdot\right]$, $V(\cdot)$, and $\mathrm{Cov}\left(\cdot\right)$ operate over both i and j where appropriate.

Continuing:

$$
\begin{aligned}
V[\hat{\beta}] &\approx \frac{1}{\mathrm{E}\left[(X_{1ij}^{r})^2\right]^2} \frac{mkV[e_{ij}X_{1ij}^{r}] + m\sum_{j=1}^{k}\sum_{j'\neq j}\mathrm{Cov}\left(e_{ij}X_{1ij}^{r}, e_{ij'}D_{ij'}^{r}\right)}{n^2} \\
&= \frac{1}{\mathrm{E}\left[(X_{1ij}^{r})^2\right]^2}\left[\frac{V[e_{ij}X_{1ij}^{r}]}{n} + \frac{m\sum_{j=1}^{k}\sum_{j'\neq j}\mathrm{Cov}\left(e_{ij}X_{1ij}^{r}, e_{ij'}D_{ij'}^{r}\right)}{n^2}\right],
\end{aligned}
$$

so if the correlation between observations within clusters is positive (as is usually the case), clustering increases the variance of our regression estimates. By how much though?

Let's impose a little working assumption. Suppose that:

- for all j, j', $V(e_{ij}X_{1ij}^{r}) = V(e_{ij'}D_{ij'}^{r})$
- for all j, j', j'', j''', $\rho(e_{ij}X_{1ij}^{r}, e_{ij'}D_{ij'}^{r}) = \rho(e_{ij''}D_{ij''}^{r}, e_{ij'''}D_{ij'''}^{r}) = \rho > 0$.

Then:

$$
\begin{aligned}
V[\hat{\beta}] &\approx \frac{1}{\mathrm{E}\left[(X_{1ij}^{r})^2\right]^2}\left[\frac{V[e_{ij}X_{1ij}^{r}]}{n} + \frac{m\sum_{j=1}^{k}\sum_{j'\neq j}\mathrm{Cov}\left(e_{ij}X_{1ij}^{r}, e_{ij'}D_{ij'}^{r}\right)}{n^2}\right] \\
&= \frac{1}{\mathrm{E}\left[(X_{1ij}^{r})^2\right]^2}\left[\frac{V[e_{ij}X_{1ij}^{r}]}{n} + \frac{m\sum_{j=1}^{k}\sum_{j'\neq j}\rho V[e_{ij}X_{1ij}^{r}]}{n^2}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\mathrm{E}\left[(X_{1ij}^r)^2\right]^2}\left[\frac{V[e_{ij}X_{1ij}^r]}{n} + \frac{mk(k-1)\rho V[e_{ij}X_{1ij}^r]}{n^2}\right] \\
&= \frac{1}{\mathrm{E}\left[(X_{1ij}^r)^2\right]^2}\left[\frac{V[e_{ij}X_{1ij}^r]}{n} + \frac{(k-1)\rho V[e_{ij}X_{1ij}^r]}{n}\right] \\
&= [1+(k-1)\rho]\frac{1}{\mathrm{E}\left[(X_{1ij}^r)^2\right]^2}\frac{V[e_{ij}X_{1ij}^r]}{n} \\
&= [1+(k-1)\rho]\frac{1}{n}\frac{V[e_{ij}X_{1ij}^r]}{V[(X_{1ij}^r)]^2}.
\end{aligned}
$$

With no "intraclass" correlation ($\rho = 0$), clustering induces no variance inflation, and it's as though we have n observations. With perfect intraclass correlation ($\rho = 1$), it's as though we only have m observations (since $k/n = 1/m$).

Usually, the answer lies somewhere in between. We don't have to hypothesize this though, we can directly estimate the variance from the data.

3.3.1 Estimating Cluster-Robust Standard Errors

A natural plug-in estimator for the sampling variance is given by plugging sample analogues into:

$$
\frac{1}{\mathrm{E}\left[(X_{1ij}^r)^2\right]^2}\left[\frac{V[e_{ij}X_{1ij}^r]}{n} + \frac{m\sum_{j=1}^{k}\sum_{j'\neq j}\mathrm{Cov}\left(e_{ij}X_{1ij}^r, e_{ij'}D_{ij'}^r\right)}{n^2}\right].
$$

The generalization (for the full regression) is given by

$$
\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\begin{pmatrix}
e_{11}^2 & e_{11}e_{12} & e_{11}e_{13} & 0 & 0 & \cdots & 0 \\
e_{11}e_{12} & e_{12}^2 & e_{12}e_{13} & 0 & 0 & \cdots & 0 \\
e_{11}e_{13} & e_{12}e_{13} & e_{13} & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & e_{21}^2 & e_{21}e_{22} & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & e_{mk}^2
\end{pmatrix}\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}.
$$

This can be straightforwardly implemented using standard software packages.

A simple alternative is given by the "block bootstrap." Just run your regression by resampling entire *clusters* (i.e., keep the values inside the clusters fixed, and resample m clusters with replacement from the m clusters in your sample). You have i.i.d. sampling of clusters, so just replicate that process. This will usually* work well for most estimators. The block bootstrap in fact probably will work better than cluster robust standard errors with a small number of clusters. With a large number of clusters, the two standard error estimators will be very similar.

3.4 Nonlinearity and Dimensionality

Regression estimates the BLP of Y given $\mathbf{X}$ (and the BLP of the CEF). What if the CEF is nonlinear? I.e., what if $\mathrm{E}[Y|X] \neq a + bX$? Can we still use linear regression to estimate the CEF? The answer is yes, though implementation and interpretation are somewhat more complicated.

Let us return to an example from Chapter 1, Example 1.4.8. Let X and Y be random variables with $X \sim U(0,1)$ and $Y = 10X^2 + W$, where $W \sim N(0,1)$ and $X \perp\!\!\!\perp W$. As shown in Example 1.4.8, the CEF of Y given X is $\mathrm{E}[Y|X] = 10X^2$, and the BLP of Y given X is $g(X) = -5/3 + 10X$. Figure 3.4.1 reproduces Figure 1.3.2, plotting 1200 random draws of (X, Y) and superimposing the graphs of the CEF (in red) and the BLP (in blue).

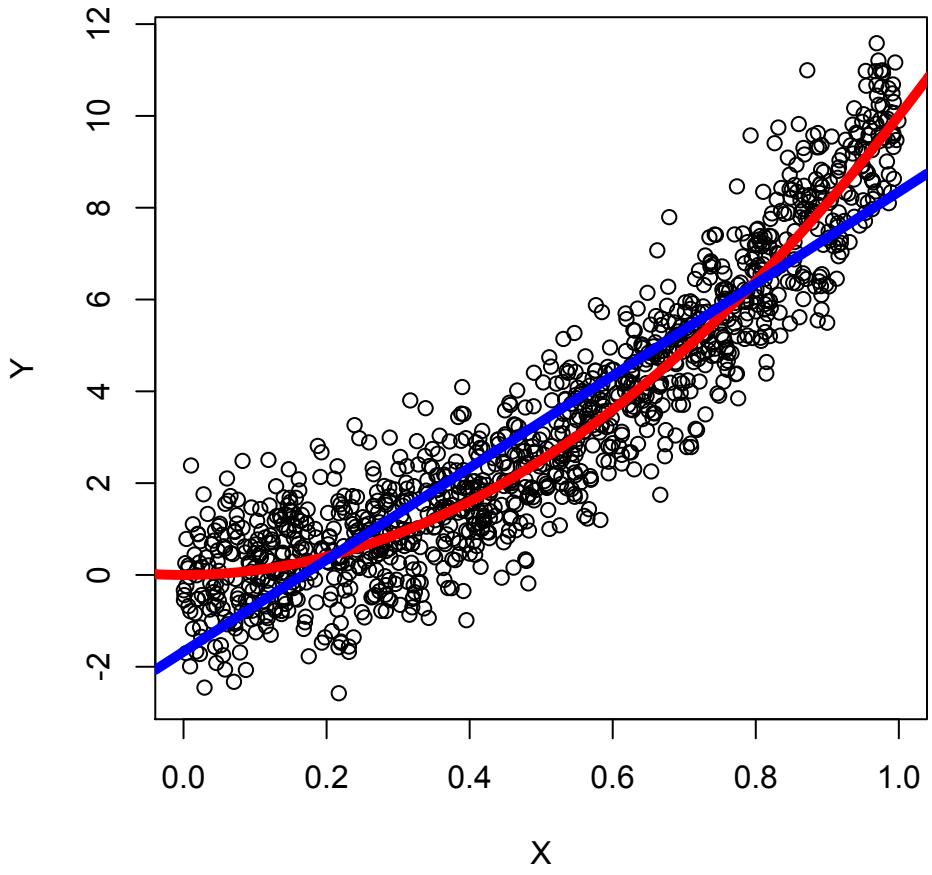


Figure 3.4.1 *Plotting the CEF and BLP*

Although the CEF is nonlinear, the BLP is a good first approximation of the CEF. Thus, linear regression, which estimates the BLP, provides a principled first approximation of the CEF. While there is no guarantee

that the BLP, and thus linear regression, will closely approximate the CEF, this is very often the case in the social and health sciences.

3.4.1 Estimation of Nonlinear CEFs

The BLP is a good, but not great, approximation to the CEF. But we can do better while still just using OLS. Continuing with the above example, suppose we now create a second explanatory variable, X^2. Then we can use linear regression to estimate:

$$\underset{(\beta_0,\beta_1,\beta_2)\in\mathbb{R}^3}{\arg\min}\ \mathrm{E}\left[\left(Y-\beta_0-\beta_1X-\beta_2X^2\right)^2\right].$$

In this example, the best linear predictor of Y given X and X^2 is the CEF.

$$g\left(X,X^2\right)=\mathrm{E}\left[Y|X,X^2\right]=\mathrm{E}\left[Y|X\right]=\beta_0+\beta_1X+\beta_2X^2=0+0X+10X^2=10X^2.$$

More generally, if the CEF of Y is linear in X and X^2 (i.e., quadratic in X),

$$\mathrm{E}\left[Y|X\right]=\beta_0+\beta_1X+\beta_2X^2,$$

then the BLP of Y given X and X^2 is the CEF, so regression of Y on X and X^2 consistently estimates the CEF. If the CEF is not linear in X and X^2, then regression of Y on X and X^2 will approximate the CEF insofar as

$$\mathrm{E}\left[Y|X\right]\approx\beta_0+\beta_1X+\beta_2X^2.$$

It is important to note that, once we introduce nonlinearity, interpreting the coefficients becomes less straightforward. With bivariate linear regression, the estimated slope of the CEF of Y with respect to X was simply $\hat{\beta}_1$. How do we compute the slope of a nonlinear function? Take the derivative.

If the CEF is linear in X and X^2, then the slope is:

$$\frac{\partial\mathrm{E}\left[Y|X\right]}{\partial X}=\frac{\partial\left(\beta_0+\beta_1X+\beta_2X^2\right)}{\partial X}=\beta_1+2\beta_2X.$$

The slope of the CEF thus depends on the value of X. So if we estimated the βs using OLS, our estimate of the slope of the CEF at $X=x$ would be $\hat{\beta}_1+2\hat{\beta}_2x$.

3.4.2 Polynomials

We can provide a more general basis for approximating the CEF using linear regression. The *Weierstrass Approximation Theorem* states that any continuous function defined on a closed interval can be uniformly approximated to arbitrary precision using a polynomial function.

Theorem 3.4.1. *Weierstrass Approximation Theorem*

Let $f:[a,b]\rightarrow\mathbb{R}$ be continuous. Then $\forall\epsilon>0$, there exists a polynomial p such that $\forall x\in[a,b]$, $|f(x)-p(x)|<\epsilon$.

We omit the proof of this theorem. We can use the Weierstrass Approximation Theorem to derive the following result.

Theorem 3.4.2. *Polynomial Approximation of the CEF*

Suppose $\mathrm{E}[Y|X=x]$ *is continuous and* $\mathrm{Supp}(X)=[a,b]$. *Then* $\forall \epsilon>0$, $\exists K\in\mathbb{N}$ *such that,* $\forall K'\geq K$,

$$\mathrm{E}\left[\left(\mathrm{E}[Y|X]-g(X,X^2,...,X^{K'})\right)^2\right]<\epsilon,$$

where $g(X,X^2,...,X^{K'})$ *is the BLP of* Y *given* X, X^2, ..., $X^{K'}$.

Proof: Let $\epsilon>0$. Then $\sqrt{\epsilon}>0$, so by the Weierstrass Approximation Theorem, there exists a polynomial p such that $\forall x\in[a,b]$,

$$|\mathrm{E}[Y|X=x]-p(x)|<\sqrt{\epsilon}.$$

Thus,

$$\mathrm{E}\left[(\mathrm{E}[Y|X]-p(X))^2\right]<\left(\sqrt{\epsilon}\right)^2=\epsilon.$$

Let K be the degree of p. The BLP of Y given X, X^2, ..., X^K is the minimum MSE predictor of $\mathrm{E}[Y|X]$ among polynomials of degree less than or equal to K, so by definition

$$\mathrm{E}\left[\left(\mathrm{E}[Y|X]-g(X,X^2,...,X^K)\right)^2\right]\leq\mathrm{E}\left[(\mathrm{E}[Y|X]-p(X))^2\right]<\epsilon,$$

where $g(X,X^2,...,X^K)$ is the BLP of Y given X, X^2, ..., X^K. Now, let $K'\geq K$. The BLP of Y given X, X^2, ..., $X^{K'}$ is the minimum MSE predictor of $\mathrm{E}[Y|X]$ among polynomials of degree less than or equal to K', so by definition

$$\mathrm{E}\left[\left(\mathrm{E}[Y|X]-g(X,X^2,...,X^{K'})\right)^2\right]\leq\mathrm{E}\left[\left(\mathrm{E}[Y|X]-g(X,X^2,...,X^K)\right)^2\right]<\epsilon.\ \square$$

Thus, as long as the CEF is continuous and the support of X is a closed interval, we can get MSE as small as we want with a polynomial BLP of sufficiently high degree.

Let's visualize this. Consider the following CEF, plotted in Figure 3.4.2:

$$\mathrm{E}[Y|X=x]=\begin{cases}0 & :x\in[-1,0]\\ x^2 & :x\in(0,1)\\ 2-x & :x\in[1,3]\end{cases}$$

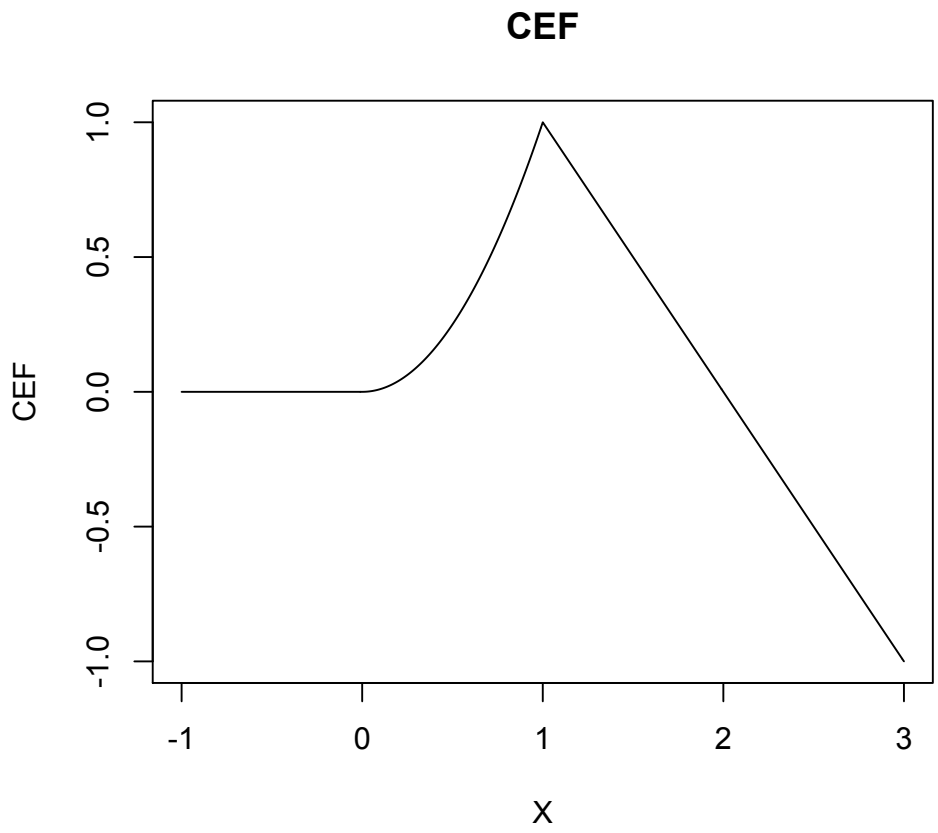


Figure 3.4.2 A Nonlinear CEF

Let's approximate this CEF with polynomials of X. Figure 3.4.3 shows the BLPs (assuming that $X \sim U(-1, 3)$) using polynomials of degrees 1, 2, 3, 4, 5, and 10.

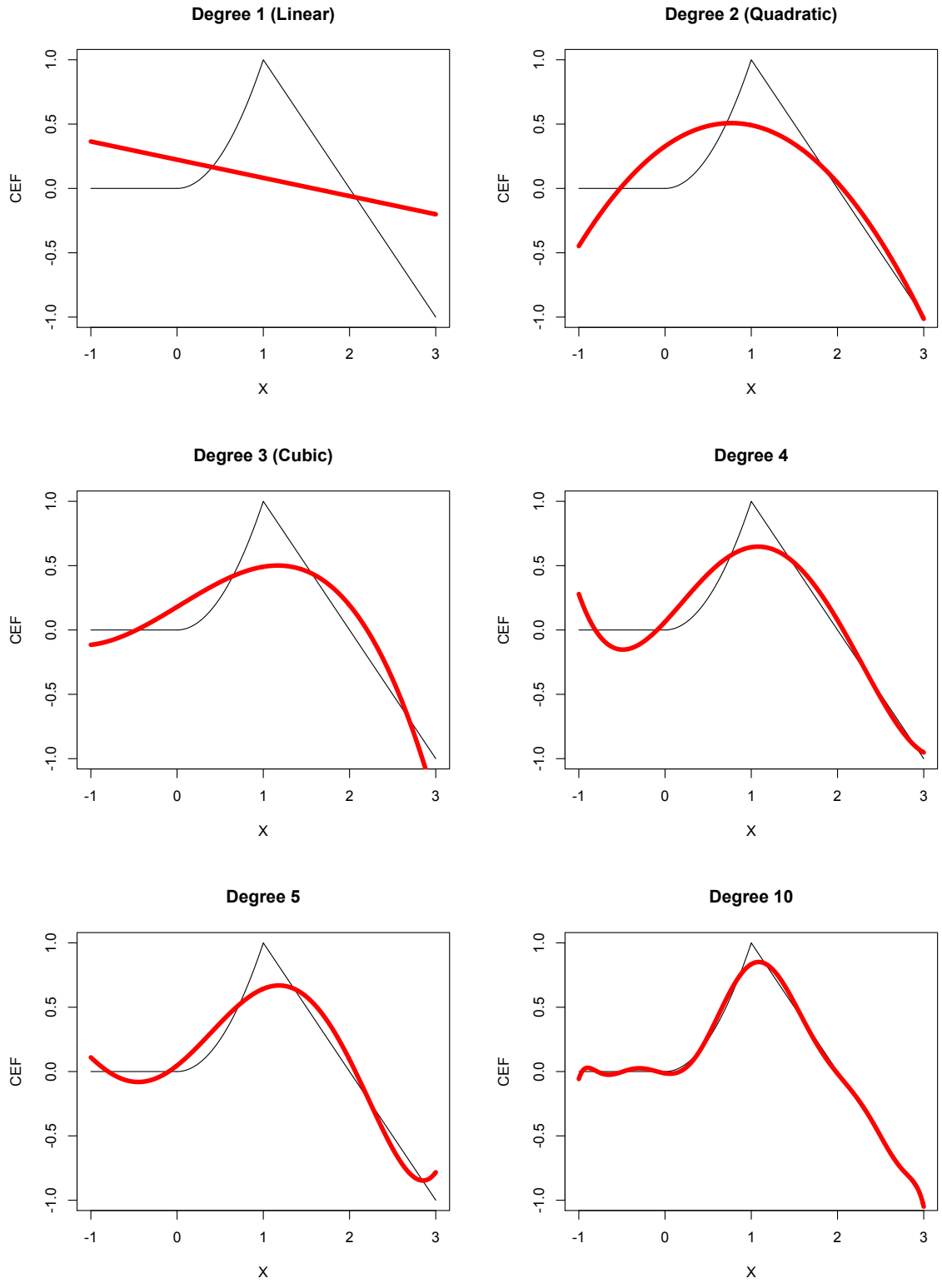


Figure 3.4.3 Polynomial Approximations of a Nonlinear CEF

This is *not* to suggest that it's good practice to include a tenth-order polynomial in your regression. But the point stands: we can always approximate the CEF to arbitrary precision using a BLP, as long as the CEF is continuous. And since regression is consistent for the BLP, this means that with large enough n, we can always *estimate* the CEF to arbitrary precision. (More on this in Section 3.4.5).

So if you're worried about the linearity assumption of linear regression, you can relax it without much issue. Polynomials are just one way. (High-order polynomials are not a particularly good way, for rather complicated reasons.) The rest of this chapter describes some others.

3.4.3 Interactions

We have shown that, with polynomials, we can approximate any continuous CEF in one variable to arbitrary precision. This was, in a sense, a proof of concept. The same principle generalizes to the multivariate case.

What happens when the slope of the CEF with respect to X_1 depends on the value of X_2? The BLP is still the BLP, but it might fail to capture an interesting *interaction* between the variables X_1 and X_2. To formalize a bit, suppose the CEF is otherwise linear, but includes an *interaction term*:

$$\mathrm{E}\left[Y|X_1, X_2\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

Much as before, we can derive the slope of the CEF with respect to each X variable by taking partial derivatives:

$$\frac{\partial \mathrm{E}\left[Y|X_1, X_2\right]}{\partial X_1} = \frac{\partial \left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2\right)}{\partial X_1} = \beta_1 + \beta_3 X_2,$$
$$\frac{\partial \mathrm{E}\left[Y|X_1, X_2\right]}{\partial X_2} = \frac{\partial \left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2\right)}{\partial X_2} = \beta_2 + \beta_3 X_1.$$

Thus, the *conditional* slope of the CEF with respect to X_1 depends on the value of X_2, and vice-versa.

In this example (unlike in Section 3.4.2), each coefficient has a clear interpretation. If the CEF is linear in X_1, X_2, and $X_1 X_2$, then

- $\beta_0 = \mathrm{E}\left[Y|X_1 = 0, X_2 = 0\right]$.
- β_1 is the slope of the CEF with respect to X_1 when $X_2 = 0$.
- β_2 is the slope of the CEF with respect to X_2 when $X_1 = 0$.
- β_3 is how each of the slopes changes when the value of the other variable is increased by one. E.g.,
  - if X_1 moves from 0 to 1, then the slope of the CEF with respect to X_2 increases by β_3.
  - if X_2 moves from 7 to 4, then the slope of the CEF with respect to X_1 decreases by $3\beta_3$.

Of course, if $\beta_3 = 0$, then $\mathrm{E}[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and there is no interaction. All of this can be generalized for any number of X variables.

For even greater flexibility, interactions can be combined with higher-order polynomial terms.[79] E.g.,

$$\mathrm{E}[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \beta_6 X_1^2 X_1 + \beta_7 X_1 X_2^2 + \beta_8 X_1^2 X_2^2.$$

Once higher-order terms are included, the coefficients no longer have straightforward interpretations. Conditional slopes with respect to each X variable can still be computed by taking partial derivatives. Beyond that, it is usually easiest to simply compute the implied value of $\mathrm{E}[Y|X_1, X_2]$ for different values of X_1 and X_2 to see how the conditional expectation of Y changes as X_1 and X_2 change.

Again, linear regression (OLS) yields a consistent estimate of the βs for any of the above specifications. When the CEF is not linear in the variables specified, OLS consistently estimates the BLP of Y given these variables, i.e., a good principled approximation of the CEF.

3.4.4 Saturated Models

Suppose we had a single binary regressor, X. I.e., $\mathrm{Supp}(X) = \{0, 1\}$. Then the CEF of Y given X is

$$\mathrm{E}[Y|X = x] = \begin{cases} \mathrm{E}[Y|X = 0] & : x = 0 \\ \mathrm{E}[Y|X = 1] & : x = 1 \end{cases}$$

This implies that we can write the CEF as a linear function of X:

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X,$$

where

- $\beta_0 = \mathrm{E}[Y|X = 0]$,
- $\beta_1 = \mathrm{E}[Y|X = 1] - \mathrm{E}[Y|X = 0]$.

Equivalently, we can write the CEF as a linear function of X and $1 - X$ without a constant:

$$\mathrm{E}[Y|X] = \gamma_1(1 - X) + \gamma_2 X,$$

where

- $\gamma_1 = \mathrm{E}[Y|X = 0]$,
- $\gamma_2 = \mathrm{E}[Y|X = 1]$.

[79] Theorem 3.4.2 can be generalized to multiple explanatory variables: if we include enough polynomial and interaction terms, we can always approximate the CEF to arbitrary precision, provided the CEF is continuous and the support of $\mathbf{X}$ is a compact set.

When X is binary, the CEF *must* be equivalent to the BLP over the support of X. The regression estimates of the βs can be computed simply by taking conditional sample means, e.g.,

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n Y_i(1-X_i)}{\sum_{i=1}^n (1-X_i)}, \; \hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i} - \frac{\sum_{i=1}^n Y_i(1-X_i)}{\sum_{i=1}^n (1-X_i)}.$$

This idea generalizes to the multivariate case. Suppose we had two binary regressors, X_1 and X_2. Then the CEF of Y given X_1 and X_2 is

$$\mathrm{E}\left[Y|X_1 = x_1, X_2 = x_2\right] = \begin{cases} \mathrm{E}\left[Y|X_1 = 0, X_2 = 0\right] & : x_1 = 0, x_2 = 0 \\ \mathrm{E}\left[Y|X_1 = 1, X_2 = 0\right] & : x_1 = 1, x_2 = 0 \\ \mathrm{E}\left[Y|X_1 = 0, X_2 = 1\right] & : x_1 = 0, x_2 = 1 \\ \mathrm{E}\left[Y|X_1 = 1, X_2 = 1\right] & : x_1 = 1, x_2 = 1 \end{cases}$$

This implies that we can write the CEF as a linear function of X_1, X_2, and X_1X_2:

$$\mathrm{E}\left[Y|X_1, X_2\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

where

- $\beta_0 = \mathrm{E}\left[Y|X_1 = 0, X_2 = 0\right]$,
- $\beta_1 = \mathrm{E}\left[Y|X_1 = 1, X_2 = 0\right] - \mathrm{E}\left[Y|X_1 = 0, X_2 = 0\right]$,
- $\beta_2 = \mathrm{E}\left[Y|X_1 = 0, X_2 = 1\right] - \mathrm{E}\left[Y|X_1 = 0, X_2 = 0\right]$,
- $\beta_3 = \mathrm{E}\left[Y|X_1 = 1, X_2 = 1\right] - \mathrm{E}\left[Y|X_1 = 1, X_2 = 0\right]$
  $- \mathrm{E}\left[Y|X_1 = 0, X_2 = 1\right] - \mathrm{E}\left[Y|X_1 = 0, X_2 = 0\right]$.

Equivalently, we can write the CEF as a linear function of "dummy" variables for every possible outcome:

$$\begin{aligned} \mathrm{E}\left[Y|X_1, X_2\right] = \gamma_1 \mathrm{I}\left[X_1 = 0, X_2 = 0\right] + \gamma_2 \mathrm{I}\left[X_1 = 1, X_2 = 0\right] \\ + \gamma_3 \mathrm{I}\left[X_1 = 0, X_2 = 1\right] + \gamma_4 \mathrm{I}\left[X_1 = 1, X_2 = 1\right], \end{aligned}$$

where

- $\gamma_1 = \mathrm{E}\left[Y|X_1 = 0, X_2 = 0\right]$,
- $\gamma_2 = \mathrm{E}\left[Y|X_1 = 1, X_2 = 0\right]$,
- $\gamma_3 = \mathrm{E}\left[Y|X_1 = 0, X_2 = 1\right]$,
- $\gamma_4 = \mathrm{E}\left[Y|X_1 = 1, X_2 = 1\right]$.

These are cases of *saturated models*: when every unique combination of values of the explanatory variables is included in a fully flexible way. This means including dummy variables for every possible value of each variable, and all possible (including multiway) interactions. Or just dummy variables for each unique combination, and excluding the intercept. Note that this is only possible with *discrete* explanatory variables and that, in such cases, the CEF is, by definition, the BLP over the support of the explanatory variables.

3.4.5 Sieve Estimation

We might be willing accept restrictive specifications with small n, keeping in mind that if we had larger n, we would increase the complexity. This idea is encapsulated in the notion of *sieve estimation*. The term sieve estimation is something of a misnomer. It's not as though it's a different type of estimator, but rather it's a way of formalizing which estimator you would choose for any given n, and then developing the asymptotic behavior of the sequence of estimators.

Intrinsically, estimating something like a CEF is an infinite-dimensional problem. We might need an infinite number of parameters to fully characterize the CEF—it's not as though we can just assume that it follows a given functional form. All estimators are limited to having so many "moving parts" given any particular n. If we want to estimate a CEF, we have to make peace with the fact that we'll just be approximating it. But we want to know that if we had very large n, we'd get it exactly.

Then the asymptotics are defined allowing for increasing flexibility, such that as n grows, the estimator gets more and more fleixble. So long as the flexibility grows a "sufficiently slow" rate, then you can show that the estimator will be consistent. We will detail one case, without getting into the math—which is rather nasty for sieve estimation.

Suppose we wanted to estimate the CEF of Y_i given X_i using OLS with polynomials, as in Section 3.4.2. Then a sieve estimator for $\mathrm{E}\left[Y_i|X_i = x_i\right]$ might look like $\hat{\mathrm{E}}\left[Y_i|X_i = x_i\right] = \sum_{k=0}^{K_n} \beta_k x_i^k$, where

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{K_n+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \sum_{k=0}^{K_n} \beta_k X_i^k \right),$$

and $K_n = \lceil n^{1/9} \rceil$. Or, in other words, we are simply regressing Y_i on a polynomial expansion of X_i, but the degree of that polynomial depends on n. Note that, as $n \to \infty$, the degree of that polynomial $K_n \to \infty$, but at a much slower rate. (E.g., when $n = 100000$, $K_n = 4$.) So long as we have positivity and the CEF is continuous, then $\hat{\mathrm{E}}\left[Y_i|X_i = x_i\right]$ will be typically consistent across regions of the data with nonzero probability mass (by Theorem 3.4.2).

Regression "works" so long as we're willing to make it more flexible as n gets large. In applied practice, it's unlikely that you'll deviate too far from linearity.

3.4.6 Penalized Regression

The final topic is penalized regression. In short, the idea is that we want to simultaneously maximize how well are predicting the outcome while minimizing the complexity of the working model that we're using.

In the case of the penalized generalization of OLS, this is formalized by introducing a "penalty" term and minimizing the sum of the penalty and the sum of squared residuals.

The perhaps best known approach here is the *lasso* (least absolute shrinkage and selection operator, Tibshirani 1996). It's penalized linear regression, and is a way of flexibly estimating the CEF, when you have either (i) a large number of variables or (ii) a large number of interactions, polynomials—the two scenarios are functionally equivalent.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{K+1}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2 + \lambda ||\boldsymbol{\beta}||_1 \right],$$

where the ℓ^1 norm, $||\boldsymbol{\beta}||_1 = \sum_{k=1}^{K} |\beta_K|$ (or the sum of the absolute value of the coefficients), and where λ is a "penalty parameter." Note that when $\lambda = 0$, this is just OLS.

We want to keep the $\boldsymbol{\beta}$ terms close to zero, unless they really help to reduce sum of squared residuals. The lasso presents a way to use many explanatory variables to "automatically" get good predictions—including when we have more variables than observations. Interpretation of lasso estimates is as you'd expect: $\hat{\boldsymbol{\beta}}$ characterizes an estimate of the BLP, and take partial derivatives where appropriate.

Choosing λ is not easy—a great deal of the machine learning literature is devoted to figuring out how to choose λ. There is no, in general, best answer for how to penalize, but the most common approach is leave-one-out crossvalidation. In this case, we choose λ via

$$\underset{\lambda_C \in \mathbb{R}^+}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}_{-i,\lambda_C})^2 \right],$$

where $\hat{\boldsymbol{\beta}}_{-i,\lambda_C}$ is computed from the lasso procedure with unit i excluded and $\lambda = \lambda_C$. This usually has good operating characteristics. So long as we assume that, as $n \to \infty$, $\lambda \not\to \infty$ and the CEF is linear in $\mathbf{X}$ (as parameterized), then the lasso will be consistent for the CEF.

There are other penalized regression techniques—ridge regression, regression trees, penalized likelihood methods, etc. All of these operate under the same basic principles, sometimes allowing for nonlinearity in different ways. As long as you understand the intuition, there's no magic here.

3.5 Thinking about Regression and Its Generalizations

All of these methods formalize or generalize principles that we've been advocating all along. If we had large n, we'd use a very flexible tool to approximate the feature of interest. With small n, we're willing to be a bit more restrictive. (I.e., using OLS to find the BLP, as a good approximation to the CEF.)

A final note: if what you're interested in is the average slope of the CEF (i.e., average marginal "effect") with respect to one variable, then—as an empirical matter—you will almost always get the same substantive answer from whatever fancy method you use as you would have if you had used OLS, even without interactions or polynomials.

Every time someone proposes a method that yields a radically different answer in characterizing the CEF of a real dataset than what OLS produces, this method is usually proven to be flawed in some way. Do not be fooled by statistical razzle-dazzle—the most basic methods work pretty darn well.

4 Missing Data

Don't play what's there; play what's not there.

— MILES DAVIS

In the preceding chapters, we have shown that, if we observe a sufficient number of i.i.d. draws of a random vector, we can estimate any feature of its distribution to arbitrary precision. Given any finite number of such observations, we can both estimate these features and, furthermore, estimate our degree of uncertainty about these estimates.

If we could always observe i.i.d. draws of any random vector whose distribution we might care about, this book could have thus concluded with Chapter 3. Unfortunately, we often want to learn about the distributions of random vectors that are *not* directly observable. What can we learn about the distribution of an *unobservable* random vector, given full knowledge of the distribution of a related *observable* random vector? This question is the core of statistical *identification*. The answer will invariably depend on what assumptions we are willing to make about the relationship between the observable and unobservable distributions.

In this chapter, we illustrate the basic concepts of identification by considering a simple but common problem: missing data. These concepts, as we shall see, all have direct analogues in the area of causal inference.

4.1 Identification with Missing Data

What can we learn about the distribution of a random variable of interest from a random sample when some values in the sample are missing? For example, if we ran a survey asking people who they voted for, some people might not respond to the vote choice question. We want to know the population distribution of outcomes for this question, but we don't observe the answers for some people.

Suppose that we are interested in estimating the expected value (or the full distribution) of some random variable Y_i.[80] However, we do not observe Y_i for any unit in our sample that does not respond. Formally, let R_i be an indicator for whether or not the outcome Y_i for unit i is observed (i.e., $R_i = 1$ if unit i responded and 0 otherwise), and let Y_i^* denote the censored version of Y_i. Then:

$$Y_i^* = Y_i R_i + (-99)(1 - R_i).^{81}$$

This equation embeds an often-overlooked assumption of *stable outcomes*, i.e., the underlying Y_i for unit i is stable and does not depend on factors like how the question was asked, who was asked, or who responded.

[80] Since we are assuming i.i.d. sampling, our inferences do not depend on the unit subscript i. I.e., learning about the random variable Y_i is equivalent to learning about the population distribution of Y_i values.

[81] The value -99 is commonly used in datasets to denote missing data.

Suppose that our inferential target is the expected value of Y_i, $\mathrm{E}[Y_i]$. We do *not* directly observe i.i.d. draws of Y_i; we only observe i.i.d. draws of the random vector (Y_i^*, R_i). Thus, without making further assumptions about the joint distribution of Y_i and R_i, we cannot directly estimate $\mathrm{E}[Y_i]$. So the question is: given full knowledge of the joint distribution of (Y_i^*, R_i), what assumptions yield what information about $\mathrm{E}[Y_i]$?

4.1.1 Bounds

Let's start with a minimal assumption. Suppose that Y_i is *bounded*, i.e., $\mathrm{Supp}(Y_i) \subseteq [a, b]$, for some $-99 < a \leq b$. Under this assumption, we can derive *sharp bounds*—i.e., bounds that cannot be improved upon without further assumptions—on $\mathrm{E}[Y_i]$. These bounds are a special case of Manski bounds.[82]

Manski bounds proceed by imputing the "worst-case scenario" for all missing data. To compute a lower bound, assume that all missing values of Y_i equal the lowest possible value, a. For an upper bound, assume that all missing values of Y_i equal the highest possible value, b.

Example 4.1.1. *Missing Data with Binary Outcomes*

Suppose that Y_i is binary, i.e., $\mathrm{Supp}(Y_i) \subseteq \{0, 1\} \subseteq [0, 1]$. We might observe the following data:

Unit	Y_i^*	R_i
1	1	1
2	-99	0
3	1	1
4	0	1
5	1	1
6	-99	0

Given the assumption of stable outcomes, this implies:

Unit	Y_i	R_i
1	1	1
2	?	0
3	1	1
4	0	1
5	1	1
6	?	0

For units 2 and 6, we only know that $Y_i \in \{0, 1\}$. To estimate a lower bound, we simply plug in 0 for these missing values:

[82] See, e.g., Manski (2003).

Unit	Y_i	R_i
1	1	1
2	0	0
3	1	1
4	0	1
5	1	1
6	0	0

Then, using the sample mean as a plug-in estimator for the expected value, our estimated lower bound for $\mathrm{E}[Y_i]$ is $1/2$. Likewise, to estimate an upper bound, we plug in 1 for the missing values:

Unit	Y_i	R_i
1	1	1
2	1	0
3	1	1
4	0	1
5	1	1
6	1	0

Thus, our estimated upper bound for $\mathrm{E}[Y_i]$ is $5/6$.

More generally, we can write down exact sharp bounds for $\mathrm{E}[Y_i]$ in terms of the joint distribution of (Y_i^*, R_i), which can, in theory, be estimated to arbitrary precision.

Theorem 4.1.1. *Sharp Bounds for the Expected Value*

Let Y_i and R_i be random variables with $\mathrm{Supp}(Y_i) \subseteq [a, b]$ and $\mathrm{Supp}(R_i) = \{0, 1\}$, and let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$. Then

$$\begin{aligned}\mathrm{E}[Y_i] \in \Big[&\mathrm{E}[Y_i^*|R_i = 1]\Pr(R_i = 1) + a\Pr(R_i = 0),\\ &\mathrm{E}[Y_i^*|R_i = 1]\Pr(R_i = 1) + b\Pr(R_i = 0)\Big].\end{aligned}$$

Proof: By the Law of Iterated Expectations,

$$\mathrm{E}[Y_i] = \mathrm{E}[Y_i|R_i = 1]\Pr(R_i = 1) + \mathrm{E}[Y_i|R_i = 0]\Pr(R_i = 0).$$

Then since $Y_i^* = Y_i R_i + (-99)(1 - R_i)$,

$$\begin{aligned}\mathrm{E}[Y_i^*|R_i = 1] &= \mathrm{E}[Y_i R_i + (-99)(1 - R_i)|R_i = 1]\\ &= \mathrm{E}[Y_i \cdot 1 + (-99) \cdot 0|R_i = 1]\\ &= \mathrm{E}[Y_i|R_i = 1].\end{aligned}$$

Thus,

$$\mathrm{E}\left[Y_i\right] = \mathrm{E}\left[Y_i^*|R_i=1\right]\Pr(R_i=1) + \mathrm{E}\left[Y_i|R_i=0\right]\Pr(R_i=0).$$

Since $\mathrm{Supp}\left(Y_i\right) \subseteq [a,b]$, $\mathrm{E}\left[Y_i|R_i=0\right] \geq a$, and therefore

$$\begin{aligned}\mathrm{E}\left[Y_i\right] &= \mathrm{E}\left[Y_i^*|R_i=1\right]\Pr(R_i=1) + \mathrm{E}\left[Y_i|R_i=0\right]\Pr(R_i=0) \\ &\geq \mathrm{E}\left[Y_i^*|R_i=1\right]\Pr(R_i=1) + a\Pr(R_i=0).\end{aligned}$$

Likewise, since $\mathrm{Supp}\left(Y_i\right) \subseteq [a,b]$, $\mathrm{E}\left[Y_i|R_i=0\right] \leq b$, and therefore

$$\begin{aligned}\mathrm{E}\left[Y_i\right] &= \mathrm{E}\left[Y_i^*|R_i=1\right]\Pr(R_i=1) + \mathrm{E}\left[Y_i|R_i=0\right]\Pr(R_i=0) \\ &\leq \mathrm{E}\left[Y_i^*|R_i=1\right]\Pr(R_i=1) + b\Pr(R_i=0).\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{E}\left[Y_i\right] \in \Big[&\mathrm{E}\left[Y_i^*|R_i=1\right]\Pr(R_i=1) + a\Pr(R_i=0), \\ &\mathrm{E}\left[Y_i^*|R_i=1\right]\Pr(R_i=1) + b\Pr(R_i=0)\Big]. \ \square\end{aligned}$$

A natural estimator here substitutes sample means for expected values. Confidence intervals for each bound can be computed using the bootstrap.

Theorem 4.1.2. *Estimating Sharp Bounds for the Expected Value*

Let Y_i and R_i be random variables with $\mathrm{Supp}\left(Y_i\right) \subseteq [a,b]$ and $\mathrm{Supp}\left(R_i\right) = \{0,1\}$, and let $Y_i^* = Y_iR_i + (-99)(1-R_i)$. Then, given n i.i.d. observations of (Y_i^*, R_i), the plug-in estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i^*R_i + a(1-R_i)\right]$$

is unbiased and consistent for the lower bound for $\mathrm{E}\left[Y_i\right]$. Likewise, the plug-in estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i^*R_i + b(1-R_i)\right]$$

is unbiased and consistent for the upper bound for $\mathrm{E}\left[Y_i\right]$.

Proof: Let $Y_i^L = Y_i^*R_i + a(1-R_i)$, so that

$$\overline{Y_i^L} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i^*R_i + a(1-R_i)\right].$$

By Theorem 2.1.1 and the WLLN, $\overline{Y_i^L}$ is unbiased and consistent for $\mathrm{E}\left[Y_i^L\right]$. And by the Law of Iterated Expectations,

$$\begin{aligned}\mathrm{E}\left[Y_i^L\right] &= \mathrm{E}\left[Y_i^L|R_i=1\right]\Pr(R_i=1) + \mathrm{E}\left[Y_i^L|R_i=0\right]\Pr(R_i=0) \\ &= \mathrm{E}\left[Y_i^*R_i + a(1-R_i)|R_i=1\right]\Pr(R_i=1) \\ &+ \mathrm{E}\left[Y_i^*R_i + a(1-R_i)|R_i=0\right]\Pr(R_i=0) \\ &= \mathrm{E}\left[Y_i^*\cdot 1 + a\cdot 0|R_i=1\right]\Pr(R_i=1) + \mathrm{E}\left[Y_i^*\cdot 0 + a\cdot 1|R_i=0\right]\Pr(R_i=0) \\ &= \mathrm{E}\left[Y_i^*|R_i=1\right]\Pr(R_i=1) + a\Pr(R_i=0).\end{aligned}$$

Thus, $\overline{Y_i^L}$ is unbiased and consistent for $\mathrm{E}\left[Y_i^* | R_i = 1\right] \Pr(R_i = 1) + a \Pr(R_i = 0)$.

Similarly, let $Y_i^U = Y_i^* R_i + b(1 - R_i)$, so that

$$\overline{Y_i^U} = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i^* R_i + b(1 - R_i) \right].$$

Then by the same logic, it follows that $\overline{Y_i^U}$ is unbiased and consistent for $\mathrm{E}\left[Y_i^* | R_i = 1\right] \Pr(R_i = 1) + b \Pr(R_i = 0)$. $\square$

Note that these plug-in estimators will reproduce our bounds estimate of $[1/2, 5/6]$ from Example 4.1.1.

Without further assumptions, these bounds are the absolute best we can do. Thus, under these minimal assumptions, $\mathrm{E}\left[Y_i\right]$ is *partially identified* (or *set identified*) as opposed to *point identified*. I.e., even with full knowledge of the joint distribution of (Y_i^*, R_i), we can do no better than the above bounds. We are left with an interval of plausible values, none of which can be logically ruled out given the assumptions that we have imposed. We cannot identify the exact value of $\mathrm{E}\left[Y_i\right]$.

Adding more assumptions inherently reduces the credibility of our estimate. With stronger assumptions, we might be able to narrow the interval, perhaps even to a point. But the cost is that our estimate becomes less believable. Manski (2003) refers to this as the "Law of Decreasing Credibility."

Sometimes, we should not feel comfortable strengthening our assumptions. Bounded support of outcomes may be the only assumption that we're willing to impose given what we know—or more to the point, what we *don't* know—about how our data were generated. In this case, if the bounds are not good enough, we need to go out and actually be *scientists*. Further information about the generative process—in this case, the determinants of non-response—is necessary to justify the stronger (i.e., more restrictive) assumptions required obtain a more precise estimate.

4.1.2 Missing Completely at Random

Let's consider one of these stronger assumptions. Suppose that the data are *missing completely at random*. This is generally considered to be strongest of all the nonparametric assumptions that we could impose.

Definition 4.1.1. *Missing Completely at Random (MCAR)*

Let Y_i and R_i be random variables with $\mathrm{Supp}\left(R_i\right) = \{0, 1\}$, and let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$. Then Y_i is missing completely at random if the following conditions hold:

- $Y_i \perp\!\!\!\perp R_i$. *(Independence of outcome and response)*
- $\Pr(R_i = 1) > 0$. *(Nonzero probability of response)*

In other words, the distribution of outcomes is exactly the same for people who respond as for people who don't respond. Under this assumption, $\mathrm{E}\left[Y_i\right]$ is *point identified*, i.e., we can write an exact expression for $\mathrm{E}\left[Y_i\right]$ in terms of the joint distribution of observables, (Y_i^*, R_i).

Theorem 4.1.3. *Expected Value under MCAR*

Let Y_i and R_i be random variables with $\text{Supp}(R_i) = \{0, 1\}$*, and let* $Y_i^* = Y_i R_i + (-99)(1 - R_i)$*. Then if Y_i is missing completely at random,*

$$\text{E}[Y_i] = \text{E}[Y_i^* | R_i = 1].$$

Proof: By independence (see Theorem 1.4.26),

$$\text{E}[Y_i] = \text{E}[Y_i | R_i = 1].$$

And by stable outcomes,

$$\text{E}[Y_i | R_i = 1] = \text{E}[Y_i^* | R_i = 1].$$

Thus,

$$\text{E}[Y_i] = \text{E}[Y_i^* | R_i = 1]. \ \square$$

Assuming MCAR thus makes it extremely easy to estimate the population mean when we have missing data. The plug-in estimator is just the sample mean of the non-missing values:

$$\hat{\text{E}}[Y_i] = \hat{\text{E}}[Y_i^* | R_i = 1] = \frac{\sum_{i=1}^n Y_i^* R_i}{\sum_{i=1}^n R_i}.$$

E.g., in Example 4.1.1, our estimate of $\text{E}[Y_i]$ assuming MCAR would be $3/4$. Our ability to draw inferences with missing data thus entirely depends on the strength of the assumptions that we're willing to impose.

One way to think about MCAR is as follows: given a large enough n, we can impute the outcomes for non-responders using the sample mean for responders. Again, consider Example 4.1.1. We have:

Unit	Y_i	R_i
1	1	1
2	?	0
3	1	1
4	0	1
5	1	1
6	?	0

Assuming MCAR, $\text{E}[Y_i | R_i = 0] = \text{E}[Y_i]$, which we can estimate using the above plug-in estimator. Thus, we can impute the missing values as follows:

Unit	Y_i	R_i
1	1	1
2	$\hat{\text{E}}[Y_i] = 3/4$	0
3	1	1
4	0	1
5	1	1
6	$\hat{\text{E}}[Y_i] = 3/4$	0

Then the sample mean with the missing values imputed is $3/4$. Of course, under the MCAR assumption, imputation is essentially a pointless extra step, since we're just using $\hat{\mathrm{E}}[Y_i]$ to impute the missing values in order to get $\hat{\mathrm{E}}[Y_i]$ again. But the utility of this way of thinking about estimating expectations with missing data will become clear in Section 4.2.

4.1.3 Adding Covariates

If we have additional information on each unit, we can impose a weaker version of the MCAR assumption. This assumption is known simply as *missing at random (MAR)* or *ignorability*. MAR is just like MCAR except *conditional on covariates*. (So it's not missing *completely* at random, just missing at random once we condition on covariates.)

Formally, suppose that, for each unit i, we observe a vector of covariates $\mathbf{X}_i$. Importantly, we observe $\mathbf{X}_i$ even when $R_i = 0$. I.e., we know the covariate values for all units; we only have missing data on the outcome variable Y_i. Now the random vector of observables is $(Y_i^*, R_i, \mathbf{X}_i)$.

Then the ignorability assumption looks just like MCAR except that everything is conditional on $\mathbf{X}_i$.

Definition 4.1.2. *Missing at Random (MAR)*

Let Y_i and R_i be random variables with $\mathrm{Supp}(R_i) = \{0, 1\}$, let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$, and let $\mathbf{X}_i$ be a random vector. Then Y_i is missing at random (or ignorable) conditional on $\mathbf{X}_i$ if the following conditions hold:

- $Y_i \perp\!\!\!\perp R_i | \mathbf{X}_i = \mathbf{x}_i$. *(Independence of outcome and response conditional on $\mathbf{X}_i$)*[83]
- $\Pr(R_i = 1 | \mathbf{X}_i) > 0$. *(Nonzero probability of response conditional on $\mathbf{X}_i$)*[84]

The first of these conditions is known as a *conditional independence assumption (CIA)*. We will see other conditional independence assumptions in later chapters. In words, the CIA says that, among units with the same measured background characteristics, the types that respond and the types that don't respond are exactly the same in terms of their distribution of Y_i values.

Almost every common method of dealing with missing data in applied practice (multiple imputation, reweighting, etc.) depends on some variant of the ignorability assumption.

How does ignorability facilitate identification of the distribution of Y_i? Though it is weaker than MCAR, this assumption still allows us to point identify $\mathrm{E}[Y_i]$.

[83] Formally, conditional independence is defined as follows: for random variables X, Y, and Z with joint PMF/PDF $f(x, y, z)$, $Y \perp\!\!\!\perp Z | X$ if, $\forall (x, y, z) \in \mathbb{R}^3$ with $x \in \mathrm{Supp}(X)$,

$$f_{(Y,Z)|X}((y, z)|x) = f_{Y|X}(y|x) f_{Z|X}(z|x).$$

If $Y \perp\!\!\!\perp Z | X$, all of the implications of independence derived in Chapter 1 hold conditional on X, e.g. $\mathrm{E}[Y|X] = \mathrm{E}[Y|Z, X]$.

[84] The notation $\Pr(\cdot|X)$ is defined analogously to $\mathrm{E}[\cdot|X]$ and $V(\cdot|X)$. E.g., if we define the function $P_{\{Y=y\}}(x) = \Pr(Y = y|X = x)$, then $\Pr(Y = y|X)$ denotes the random variable $Z = P_{\{Y=y\}}(X)$. Hence, the statement that $\Pr(R_i = 1|\mathbf{X}_i) > 0$ is equivalent to: $\forall \mathbf{x}_i \in \mathrm{Supp}(\mathbf{X}_i), \Pr(R_i = 1|\mathbf{X}_i = \mathbf{x}_i) > 0$.

Theorem 4.1.4. *Expected Value under Ignorability*

Let Y_i and R_i be random variables with $\text{Supp}\,(R_i) = \{0, 1\}$*, let* $Y_i^* = Y_i R_i + (-99)(1 - R_i)$*, and let* $\mathbf{X}_i$ *be a discrete random vector.*[85] *Then if Y_i is missing at random conditional on $\mathbf{X}_i$)*

$$\text{E}\,[Y_i] = \sum_{\mathbf{x}_i} \text{E}\,[Y_i^* | R_i = 1, \mathbf{X}_i = \mathbf{x}_i]\,\Pr(\mathbf{X}_i = \mathbf{x}_i).$$

Proof: By the Law of Iterated Expectations,

$$\text{E}\,[Y_i] = \text{E}_{\,\mathbf{X}_i}\,[\text{E}\,[Y_i | \mathbf{X}_i]] = \sum_{\mathbf{x}_i} \text{E}\,[Y_i | \mathbf{X}_i = \mathbf{x}_i]\,\Pr(\mathbf{X}_i = \mathbf{x}_i).$$

By MAR (and stable outcomes),

$$\text{E}\,[Y_i | \mathbf{X}_i] = \text{E}\,[Y_i | R_i = 1, \mathbf{X}_i] = \text{E}\,[Y_i^* | R_i = 1, \mathbf{X}_i].$$

Thus,

$$\text{E}\,[Y_i] = \sum_{\mathbf{x}_i} \text{E}\,[Y_i^* | R_i = 1, \mathbf{X}_i = \mathbf{x}_i]\,\Pr(\mathbf{X}_i = \mathbf{x}_i). \;\square$$

Since we can observe the random vector $(Y_i^*, R_i, \mathbf{X}_i)$, we can point identify $\text{E}\,[Y_i^* | R_i = 1, \mathbf{X}_i = \mathbf{x}_i]$ and $\Pr(\mathbf{X}_i = \mathbf{x}_i)$. Thus, under ignorability, $\text{E}\,[Y_i]$ is point identified.

Furthermore, note what the second line of the above proof implies: the CEF of Y_i given $\mathbf{X}_i$ is point identified under ignorability.

Theorem 4.1.5. *The CEF under Ignorability*

Let Y_i and R_i be random variables with $\text{Supp}\,(R_i) = \{0, 1\}$*, let* $Y_i^* = Y_i R_i + (-99)(1 - R_i)$*, and let* $\mathbf{X}_i$ *be a random vector. Then if Y_i is missing at random conditional on X_i,*

$$\text{E}\,[Y_i | \mathbf{X}_i] = \text{E}\,[Y_i^* | R_i = 1, \mathbf{X}_i].$$

Thus, under ignorability, the CEF of observable outcomes is identical to the CEF of all outcomes.

4.2 Estimation with Missing Data

We have shown that, under the MCAR assumption, estimating $\text{E}\,[Y_i]$ is extremely straightforward: just use the sample mean of the observed outcomes. Under MAR, the intuition is essentially the same, but estimation becomes a bit more complicated. In this section, we discuss several ways to implement the MAR assumption to obtain an estimate.

[85] We assume here that $\mathbf{X}_i$ is discrete to ease exposition. This assumption is not necessary, but it lets us work with PMFs instead of PDFs. The continuous case is analogous. Recall that we are writing $\sum_{\mathbf{x}_i}$ as a shorthand for $\sum_{\mathbf{x}_i \in \text{Supp}\,(\mathbf{X}_i)}$.

4.2.1 Plug-in Estimation

If we have discrete covariates, then under the ignorability assumption, we can consistently estimate $\mathrm{E}[Y_i]$ with a simple plug-in estimator:

$$\begin{aligned}\hat{\mathrm{E}}[Y_i] &= \sum_{\mathbf{x}_i \in \widehat{\mathrm{Supp}}(\mathbf{X}_i)} \hat{\mathrm{E}}[Y_i^* | R_i = 1, \mathbf{X}_i = \mathbf{x}_i] \widehat{\Pr}(\mathbf{X}_i = \mathbf{x}_i) \\ &= \sum_{\mathbf{x}_i \in \widehat{\mathrm{Supp}}(\mathbf{X}_i)} \frac{\sum_{i=1}^n Y_i^* R_i \cdot \mathrm{I}(\mathbf{X}_i = \mathbf{x}_i)}{\sum_{i=1}^n R_i \cdot \mathrm{I}(\mathbf{X}_i = \mathbf{x}_i)} \cdot \frac{\sum_{i=1}^n \mathrm{I}(\mathbf{X}_i = \mathbf{x}_i)}{n},\end{aligned}$$

where $\mathrm{I}(\cdot)$ is the indicator function and $\widehat{\mathrm{Supp}}(\mathbf{X}_i)$ is the set of all observed values of $\mathbf{X}_i$ (i.e., the plug-in estimator for the support of $\mathbf{X}_i$).[86] Note that this requires that we have at least one unit with $R_i = 1$ and $\mathbf{X}_i = \mathbf{x}_i$ for every covariate profile $\mathbf{x}_i \in \widehat{\mathrm{Supp}}(\mathbf{X}_i)$, otherwise we'll have a zero in the denominator.

Consider again Example 4.1.1, but now assume that we also observe a binary covariate X_i for every unit.

Unit	Y_i	R_i	X_i
1	1	1	0
2	?	0	0
3	1	1	0
4	0	1	0
5	1	1	1
6	?	0	1

Assuming MAR, the above plug-in estimator yields the estimate

$$\hat{\mathrm{E}}[Y_i] = \hat{\mathrm{E}}[Y_i^* | R_i = 1, X_i = 0] \widehat{\Pr}(X_i = 0) + \hat{\mathrm{E}}[Y_i^* | R_i = 1, X_i = 1] \widehat{\Pr}(X_i = 1) = \frac{2}{3} \cdot \frac{4}{6} + 1 \cdot \frac{2}{6} = \frac{7}{9}.$$

Note that this estimate differs from the MCAR-based estimate of $3/4$.

Like MCAR, we can think of the MAR assumption as allowing us to impute the outcomes for non-responders, only now we impute each missing outcome using the sample mean of respondents *with the same covariate values*, since under ignorability, $\forall \mathbf{x}_i \in \mathrm{Supp}(\mathbf{X}_i), \mathrm{E}[Y_i | R_i = 0, \mathbf{X}_i = \mathbf{x}_i] = \mathrm{E}[Y_i | \mathbf{X}_i = \mathbf{x}_i]$. In Example 4.1.1, this yields

Unit	Y_i	R_i	X_i
1	1	1	0
2	$\hat{\mathrm{E}}[Y_i \mid X_i = 0] = 2/3$	0	0
3	1	1	0
4	0	1	0
5	1	1	1
6	$\hat{\mathrm{E}}[Y_i \mid X_i = 1] = 1$	0	1

[86] This estimator is sometimes known as the *post-stratification* estimator for missing data.

Then the sample mean with the missing values imputed is $7/9$, which is just the plug-in estimate.

Similar methods exist for filling in *joint* distributions when you also have missing values for $\mathbf{X}_i$ as well as outcomes. You just need to invoke ignorability with respect to all of the non-missing variables. When you use, e.g., multiple imputation, ignorability is the type of assumption that you are invoking.

The logic behind this plug-in estimator is also the basis for survey reweighting. Suppose that we have undersampled a certain group characterized by $\mathbf{X}_i = \mathbf{x}_i$, and suppose that we know the population distribution of $\mathbf{X}_i$. If we assume ignorability, we can obtain a consistent estimate using the above plug-in estimator.

Example 4.2.1. *Survey Reweighting*

Suppose that we conducted an Internet survey asking a large number of adult U.S. citizens (1) their height (in inches) and (2) their gender. Assume that no one lies. Our goal is to estimate the average height of adult Americans. The problem is: people who answer Internet surveys may be unusual, both in ways we can measure and in ways we can't.

Suppose that we obtained results characterized by the following summary statistics:

$$\begin{aligned}\hat{\mathrm{E}}\left[Y_i^*|R_i=1,X_i=0\right]&=64, \quad \widehat{\Pr}(X_i=0|R_i=1)=\tfrac{3}{10},\\ \hat{\mathrm{E}}\left[Y_i^*|R_i=1,X_i=1\right]&=70, \quad \widehat{\Pr}(X_i=1|R_i=1)=\tfrac{7}{10},\end{aligned}$$

where X_i is an indicator variable for gender.[87] The unadjusted estimate of the national average height is

$$\begin{aligned}\hat{\mathrm{E}}\left[Y_i\right] &= \hat{\mathrm{E}}\left[Y_i^*|R_i=1,X_i=0\right]\widehat{\Pr}(X_i=0|R_i=1)+\hat{\mathrm{E}}\left[Y_i^*|R_i=1,X_i=1\right]\widehat{\Pr}(X_i=1|R_i=1)\\ &= 64\cdot\frac{3}{10}+70\cdot\frac{7}{10}\\ &= 68.2 \text{ inches.}\end{aligned}$$

This is the estimate we would obtain if we used a plug-in estimator after assuming MCAR, i.e., if we simply used the sample mean as our estimate. But we can do better. Assume that the adult U.S. population is 50% women, so we know that $\Pr(X_i=0)=\Pr(X_i=1)=1/2$. Then under ignorability with respect to X_i:

$$\begin{aligned}\mathrm{E}\left[Y_i\right] &= \sum_{x_i}\mathrm{E}\left[Y_i|X_i=x_i\right]\Pr(X_i=x_i)\\ &= \sum_{x_i}\mathrm{E}\left[Y_i^*|R_i=1,X_i=x_i\right]\Pr(X_i=x_i)\\ &= \mathrm{E}\left[Y_i^*|R_i=1,X_i=0\right]\cdot\frac{1}{2}+\mathrm{E}\left[Y_i^*|R_i=1,X_i=1\right]\cdot\frac{1}{2}.\end{aligned}$$

Thus, the *adjusted* plug-in estimate of $\mathrm{E}\left[Y_i\right]$ is

$$\hat{\mathrm{E}}\left[Y_i\right]=\hat{\mathrm{E}}\left[Y_i^*|R_i=1,X_i=0\right]\cdot\frac{1}{2}+\hat{\mathrm{E}}\left[Y_i^*|R_i=1,X_i=1\right]\cdot\frac{1}{2}=64\cdot\frac{1}{2}+70\cdot\frac{1}{2}=67 \text{ inches.}$$

[87] By convention, "men are odd," i.e., $X_i=0$ for women and 1 for men.

It should be clear that the adjusted estimate is more reasonable than the unadjusted estimate. We know that the U.S. population is about 50% women. And we know (and our data confirm) that gender is associated with height—men are taller on average. So clearly the distribution of heights in a sample that is 70% men is likely to be unrepresentative of the distribution of heights in the general population. The adjusted estimate allows us to correct for this sampling bias.

However, if MAR conditional on gender does not hold, this adjusted plug-in estimator may still be inconsistent. There might be unobserved (or perhaps even unobservable) determinants of response that are related to height. Is there any reason why Internet survey respondents might tend to be shorter or taller on average than the general population, even conditional on gender?

Thus far we have dealt with estimation under MAR only in the special case where we have a single discrete covariate that can take on just a small number of values. When we have many covariates, or continuous covariates, simple plug-in estimation of this sort generally isn't feasible. In the remainder of this chapter, we discuss some other ways of estimating the population mean under ignorability.

4.2.2 Regression Estimation

Once again, consider Example 4.1.1, but now assume that we observe two covariates for every unit.

Unit	Y_i	R_i	X_{1i}	X_{2i}
1	1	1	0	3
2	?	0	0	7
3	1	1	0	9
4	0	1	0	5
5	1	1	1	4
6	?	0	1	3

Under ignorability, $\forall \mathbf{x}_i \in \text{Supp}\left(\mathbf{X}_i\right), \text{E}\left[Y_i | R_i = 0, \mathbf{X}_i = \mathbf{x}_i\right] = \text{E}\left[Y_i | \mathbf{X}_i = \mathbf{x}_i\right]$, so we want to fill in all of the missing values of Y_i with $\hat{\text{E}}\left[Y_i | X_{1i} = x_{1i}, X_{2i} = x_{2i}\right]$.

Let's see how we could do this using regression. Suppose we assumed that, at least to a first approximation,

$$\text{E}\left[Y_i | X_{1i}, X_{2i}\right] = \text{E}\left[Y_i^* | R_i = 1, X_{1i}, X_{2i}\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}.$$

(The first equality is implied by ignorability; the second is a functional form assumption.) Then we could estimate the coefficients with OLS and use the resulting equation to impute the missing values.

Unit	Y_i	R_i	X_{1i}	X_{2i}
1	1	1	0	3
2	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 7$	0	0	7
3	1	1	0	9
4	0	1	0	5
5	1	1	1	4
6	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 3$	0	1	3

(Remember: these are just expectations for units with the same covariate values; we can't *actually* fill in the missing outcomes.) Equivalently, we could instead impute all of the Y_i values, which will yield the same estimate for $\mathrm{E}[Y_i]$ and is more straightforward to implement.

Unit	Y_i	R_i	X_{1i}	X_{2i}
1	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 3$	1	0	3
2	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 7$	0	0	7
3	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 9$	1	0	9
4	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 5$	1	0	5
5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 4$	1	1	4
6	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 3$	0	1	3

In either case, the estimate of $\mathrm{E}[Y_i]$ is then just the sample mean with imputed values for potential outcomes. This estimator would be consistent if ignorability held and if the functional form of the CEF were in fact $\mathrm{E}[Y_i^*|R_i = 1, X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$.

When n is small, we will generally want to fit a simple approximation so as to reduce the variability of our estimates. You may ask, though: how can we relax the working assumption that $\mathrm{E}[Y_i^*|R_i = 1, X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$? Interactions and polynomials! Recall that, for any continuous CEF, corollaries to the Weierstrass Approximation Theorem (e.g., Theorem 3.4.2) guarantee that we can approximate $\mathrm{E}[Y_i^*|R_i = 1, X_{1i}, X_{2i}]$ to arbitrary precision in this way.

For example, we could instead impose the following functional form:

$$\mathrm{E}[Y_i^*|R_i = 1, X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{1i} X_{2i} + \beta_5 X_{2i}^2.$$

This less restrictive approximation is of course preferable with large n—the worst case scenario is that the extra coefficients are unnecessary. But with small n, we might not want to estimate so many coefficients, as this will result in highly imprecise estimates. We could therefore consider a sieve estimator, where the number of interactions and polynomials grows large as n increases (see Section 3.4.5). When the CEF is continuous over the support of the covariates (a usually minor technical assumption), flexible regression allows us to approximate the CEF arbitrarily well given sufficient data.

There are other ways to generate consistent nonparametric estimators of the CEF. The literature on nonparametric regression is vast. In particular, kernel methods (fitting each observation's associated expectation using only data from close observations) are popular, as are regression trees (choosing "cutpoints" in combinations of the covariates to best predict the outcome). We are not going to cover these, but the logic behind them is not too far from what you might expect. The goal is to flexibly approximate the CEF, and there are many ways to do so.

In general, to quickly produce an estimate of the expected value of Y_i from any of these nonparametric (or semiparametric) regression approaches, it is easiest to take predicted values for all observations and take the average of these predicted values. The logic of this procedure is clear from the plug-in principle:

$$\mathrm{E}[Y_i] = \mathrm{E}_{\mathbf{X}_i}[\mathrm{E}[Y_i|\mathbf{X}_i]],$$

so a natural plug-in estimator is the sample mean of predicted values,

$$\hat{\mathrm{E}}\left[Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathrm{E}}\left[Y_i|\mathbf{X}_i = \mathbf{x}_i\right],$$

where $\hat{\mathrm{E}}\left[Y_i|\mathbf{X}_i = \mathbf{x}_i\right]$ is any consistent estimator of the CEF.

In short, regression works as long as we're willing to make it more flexible as n gets large. In applied practice, it's unlikely that you'll deviate too far from linearity.

4.2.3 The Role of the Propensity Score

Before we discuss some other methods of estimating the expected value with missing data, we must first establish a fundamental result. If ignorability holds, then instead of directly conditioning on all of the covariates, it suffices to just condition on a summary measure of missingness and the covariates.

This summary measure is the *propensity score*. In the context of missing data, the propensity score is the conditional probability of response given the covariates. Note that the propensity score for any given covariate profile is simply a function of the covariates: if we know that $\mathbf{X}_i = \mathbf{x}_i$, and we know the full joint distribution of $(R_i, \mathbf{X}_i)$, then we know $\Pr(R_i = 1|\mathbf{X}_i = \mathbf{x}_i)$.

The following theorem states the key properties of the propensity score under ignorability.

Theorem 4.2.1. *Ignorability and the Propensity Score*

Let Y_i and R_i be random variables with $\mathrm{Supp}\left(R_i\right) = \{0, 1\}$, let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(R_i = 1|\mathbf{X}_i)$. Then if Y_i is missing at random conditional on $\mathbf{X}_i$,

- $R_i \perp\!\!\!\perp \mathbf{X}_i|P_i$. *(Balance conditional on P_i)*
- $Y_i \perp\!\!\!\perp R_i|P_i$. *(Independence of outcome and response conditional on P_i)*
- $\Pr(R_i = 1|P_i) > 0$. *(Nonzero probability of response conditional on P_i)*

Proof: We'll prove the first two properties and leave the third as an exercise to the reader.

Balance: Since R_i is binary, its distribution is fully governed by its mean, so

$$R_i \perp\!\!\!\perp \mathbf{X}_i|P_i \quad \Leftrightarrow \quad \mathrm{E}\left[R_i|\mathbf{X}_i, P_i\right] = \mathrm{E}\left[R_i|P_i\right].$$

The propensity score can be written as a function of $\mathbf{X}_i$: $P_i = f(\mathbf{X}_i)$. Thus,

$$\mathrm{E}\left[R_i|\mathbf{X}_i, P_i\right] = \mathrm{E}\left[R_i|\mathbf{X}_i, f(\mathbf{X}_i)\right] = \mathrm{E}\left[R_i|\mathbf{X}_i\right] = P_i.$$

And by the Law of Iterated Expectations,

$$\mathrm{E}\left[R_i|P_i\right] = \mathrm{E}\left[\mathrm{E}\left[R_i|\mathbf{X}_i\right]|P_i\right] = \mathrm{E}\left[P_i|P_i\right] = P_i.$$

Therefore,

$$\mathrm{E}\left[R_i|\mathbf{X}_i, P_i\right] = P_i = \mathrm{E}\left[R_i|P_i\right],$$

and thus $R_i \perp\!\!\!\perp \mathbf{X}_i|P_i$.

Independence of outcome and response: Since R_i is binary,

$$R_i \perp\!\!\!\perp Y_i|P_i \quad \Leftrightarrow \quad \mathrm{E}\left[R_i|Y_i, P_i\right] = \mathrm{E}\left[R_i|P_i\right].$$

By the Law of Iterated Expectations,

$$\mathrm{E}\left[R_i|Y_i, P_i\right] = \mathrm{E}\left[\mathrm{E}\left[R_i|Y_i, \mathbf{X}_i\right]|Y_i, P_i\right].$$

And by ignorability,

$$\mathrm{E}\left[R_i|Y_i, \mathbf{X}_i\right] = \mathrm{E}\left[R_i|\mathbf{X}_i\right] = P_i.$$

So by substitution,

$$\mathrm{E}\left[R_i|Y_i, P_i\right] = \mathrm{E}\left[P_i|Y_i, P_i\right] = P_i = \mathrm{E}\left[R_i|P_i\right],$$

and thus $R_i \perp\!\!\!\perp Y_i|P_i$. $\square$

In short, Theorem 4.2.1 says that conditioning on the propensity score is equivalent to conditioning on all of the covariates.

It is important to note that balance on $\mathbf{X}_i$ is a *consequence* of conditioning on the propensity score. I.e., after you have conditioned on the propensity score, the conditional distribution of covariates for observed units will be the same as the conditional distribution of covariates for missing units. In this context, a *balance test* (a statistical test of equality of covariate distributions) merely tests the null hypothesis that we have successfully conditioned on the propensity score. If the distributions of covariates, conditional on the propensity score, for observed and unobserved units are significantly different, then we have done something wrong. However, balance does *not* imply that ignorability holds. Balance on observable characteristics does not imply balance on unobservable characteristics. Without further assumptions, there exists no general test of ignorability.

We have shown that, in order to obtain consistent estimates of the population mean, we do not need to condition on all of the covariates as long as we condition on the propensity score. This result is thought to be useful since it might be difficult to (nonparametrically) condition on a large number of variables (unless we use a simple approximation like OLS). But there is a problem here: with real data, the propensity score must be *estimated*.

How would we estimate the propensity score? By performing some sort of regression of R_i on all of the covariates. We have thus simply shifted the problem from estimating $\mathrm{E}\left[Y_i^*|R_i = 1, \mathbf{X}_i\right]$ to estimating $\mathrm{E}\left[Y_i^*|R_i = 1, P_i\right]$ and $P_i = \mathrm{E}\left[R_i|\mathbf{X}_i\right]$. Nevertheless, this is what must be done when working with the propensity score.

4.2.4 Hot Deck Imputation

Another method of imputing missing outcomes is *hot deck imputation*. Under ignorability, hot deck imputation will typically yield a consistent (but otherwise pretty bad) estimator of the population mean.

There are many types of hot deck imputation, but we will discuss just one type: one-to-one, nearest-neighbor, with-replacement, propensity score hot deck imputation. For the moment, we'll ignore the need to estimate the propensity score and just assume that it is known.

Example 4.2.2. *Imputation with the Propensity Score*

Suppose that we had the following data:

Unit	Y_i	R_i	P_i
1	2	1	0.33
2	?	0	0.14
3	3	1	0.73
4	10	1	0.35
5	2	1	0.78
6	?	0	0.70

Assume that ignorability holds and that our goal is to estimate $\mathrm{E}[Y_i]$. By Theorem 4.2.1, ignorability implies that $\mathrm{E}[Y_i|R_i = 0, P_i] = \mathrm{E}[Y_i|P_i]$, so we want to impute the missing outcomes with estimates of $\mathrm{E}[Y_i|P_i = p_i]$:

Unit	Y_i	R_i	P_i
1	2	1	0.33
2	$\hat{\mathrm{E}}[Y_i\|P_i = 0.14]$	0	0.14
3	3	1	0.73
4	10	1	0.35
5	2	1	0.78
6	$\hat{\mathrm{E}}[Y_i\|P_i = 0.70]$	0	0.70

Then, if our estimator of $\mathrm{E}[Y_i|P_i = p_i]$ is consistent, the sample mean with the missing values imputed will be a consistent estimator of the population mean.

There are many imputation-type estimators for missing outcomes that exploit ignorability. (Regression is itself an imputation estimator.) The procedure for (one-to-one, nearest-neighbor) hot deck imputation is perhaps the simplest one possible: for each missing unit, simply find the observed unit that is *closest* on P_i and use that unit's outcome to fill in the missing outcome.[88] In this example, this yields:

[88] The name "hot deck imputation" derives from the era when data were stored on punch cards. To impute missing outcomes, one would use an observed value from the same dataset—a stack of cards—which was still hot because it was currently being processed.

Unit	Y_i	R_i	P_i
1	2	1	0.33
2	2	0	0.14
3	3	1	0.73
4	10	1	0.35
5	2	1	0.78
6	3	0	0.70

So the hot deck imputation estimate of $\mathrm{E}[Y_i]$ is $^{22}/_6 = {}^{11}/_3 \approx 3.67$.

Why is one-to-one, nearest-neighbor hot deck imputation not such a great idea, even when we assume ignorability? Why would you not want to just use the outcome for the closest observed unit on P_i to estimate a missing outcome? Intuitively, hot deck imputation only uses a small amount of the information available to impute the missing values. In the above example, we used $Y_1 = 2$ to impute Y_2 because $P_1 = 0.33$ is closest to $P_2 = 0.14$. But $P_4 = 0.35$ is almost just as close to P_1, so $Y_4 = 10$ probably also gives us some information about $\mathrm{E}[Y_i|P_i = 0.14]$. So why would we ignore Y_4 completely and just use Y_1 as our guess for Y_2?

The problems with this method get even worse when you have to estimate the propensity score. Suppose you don't know the functional form of $\mathrm{E}[R_i|\mathbf{X}_i = \mathbf{x}_i]$. If you aren't very careful to set up a nonparametrically consistent estimator (e.g., some variant of regression with a growing number of polynomial and interaction terms), then your estimator of P_i may never converge to the right value. Then you won't estimate the population mean consistently, even under ignorability. Furthermore, even if you do have a consistent estimator of P_i, you'll still need to account for how the uncertainty of your propensity score estimates contributes to the uncertainty of your estimate of $\mathrm{E}[Y_i]$.

4.2.5 Weighting Estimators

There are better ways to use the propensity score to estimate $\mathrm{E}[Y_i]$ with missing data. *Inverse probability weighted (IPW) estimators* provide one such method. The earliest estimator of this kind was derived for the case of survey sampling where the probabilities of response are known (Horvitz and Thompson 1952).

Definition 4.2.1. *The Horvitz-Thompson (HT) Estimator for Missing Data*

Let Y_i and R_i be random variables with $\mathrm{Supp}(R_i) = \{0, 1\}$, let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(R_i = 1|\mathbf{X}_i)$, assuming that all $P_i > 0$. Then the Horvitz-Thompson estimator for $\mathrm{E}[Y_i]$ is

$$\hat{\mathrm{E}}_{HT}[Y_i] = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^* R_i}{P_i}.$$

The Horvitz-Thompson estimator simply weights every observed Y_i by the inverse of the propensity score and takes the sample mean (with missing values effectively set equal to 0). The following theorem states the key properties of this estimator.

Theorem 4.2.2. *Properties of the HT Estimator for Missing Data*

Let Y_i and R_i be random variables with $\text{Supp}\,(R_i) = \{0, 1\}$*, let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(R_i = 1|\mathbf{X}_i)$. Then if Y_i is missing at random conditional on $\mathbf{X}_i$, the Horvitz-Thompson estimator $\hat{\text{E}}_{HT}[Y_i]$ is unbiased and consistent for* $\text{E}\,[Y_i]$.

Proof: By definition,

$$\text{E}\left[\frac{Y_i^* R_i}{P_i}\middle|\mathbf{X}_i\right] = \text{E}\left[Y_i^* \cdot \frac{R_i}{\Pr(R_i = 1|\mathbf{X}_i)}\middle|\mathbf{X}_i\right].$$

And by stable outcomes,

$$\text{E}\left[Y_i^* \cdot \frac{R_i}{\Pr(R_i = 1|\mathbf{X}_i)}\middle|\mathbf{X}_i\right] = \text{E}\left[Y_i \cdot \frac{R_i}{\Pr(R_i = 1|\mathbf{X}_i)}\middle|\mathbf{X}_i\right].$$

Then since ignorability holds, $Y_i \perp\!\!\!\perp R_i|\mathbf{X}_i$, so

$$\begin{aligned}
\text{E}\left[Y_i \cdot \frac{R_i}{\Pr(R_i = 1|\mathbf{X}_i)}\middle|\mathbf{X}_i\right] &= \text{E}\,[Y_i|\mathbf{X}_i] \cdot \text{E}\left[\frac{R_i}{\Pr(R_i = 1|\mathbf{X}_i)}\middle|\mathbf{X}_i\right] \\
&= \text{E}\,[Y_i|\mathbf{X}_i] \cdot \frac{\text{E}\,[R_i|\mathbf{X}_i]}{\Pr(R_i = 1|\mathbf{X}_i)} \\
&= \text{E}\,[Y_i|\mathbf{X}_i] \cdot \frac{\Pr(R_i = 1|\mathbf{X}_i)}{\Pr(R_i = 1|\mathbf{X}_i)} \\
&= \text{E}\,[Y_i|\mathbf{X}_i]\,.
\end{aligned}$$

Thus,

$$\text{E}\left[\frac{Y_i^* R_i}{P_i}\middle|\mathbf{X}_i\right] = \text{E}\,[Y_i|\mathbf{X}_i]\,.$$

So by the Law of Iterated Expectations,

$$\text{E}\,[Y_i] = \text{E}_{\mathbf{X}_i}[\text{E}\,[Y_i|\mathbf{X}_i]] = \text{E}_{\mathbf{X}_i}\left[\text{E}\left[\frac{Y_i^* R_i}{P_i}\middle|\mathbf{X}_i\right]\right] = \text{E}\left[\frac{Y_i^* R_i}{P_i}\right].$$

This equation suggests a simple plug-in estimator,

$$\hat{\text{E}}_{HT}[Y_i] = \hat{\text{E}}\left[\frac{Y_i^* R_i}{P_i}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^* R_i}{P_i},$$

which is unbiased and consistent by Theorem 2.1.1 and the WLLN. □

Note that the HT estimator is not necessarily unbiased if P_i is estimated, though it remains consistent so long as the estimator of P_i is consistent.

Intuitively, the logic behind IPW estimators is essentially the same as the logic of survey reweighting. Suppose that some units are very unlikely to respond, so they have small propensity scores. These units are likely to be underrepresented in our sample, so we want to up-weight them when we actually do observe them. In the case where we have a discrete and finite number of values for $\mathbf{X}_i$, the HT estimator is logically equivalent to the plug-in estimator from Section 5.2.1.

Though the HT estimator is unbiased and consistent, it has high variability in small samples. An alternative that performs better in practice is a "ratio estimator" proposed by Hajek (1971).

Definition 4.2.2. *The Hajek Estimator for Missing Data*

Let Y_i and R_i be random variables with $\text{Supp}(R_i) = \{0, 1\}$, let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(R_i = 1|\mathbf{X}_i)$. Then the Hajek estimator for $\text{E}[Y_i]$ is

$$\hat{\text{E}}_{Haj}[Y_i] = \frac{\sum_{i=1}^n \frac{Y_i^* R_i}{P_i}}{\sum_{i=1}^n \frac{R_i}{P_i}}.$$

With the Hajek estimator, we renormalize the weights to sum to n. This normalization is useful in case we draw an unusually large number of units where P_i is small. The Hajek estimator is *not* generally unbiased, but it is consistent and is usually* more efficient than the HT estimator.

Theorem 4.2.3. *Consistency of the Hajek Estimator for Missing Data*

Let Y_i and R_i be random variables with $\text{Supp}(R_i) = \{0, 1\}$, let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(R_i = 1|\mathbf{X}_i)$. Then if Y_i is missing at random conditional on $\mathbf{X}_i$, the Hajek estimator $\hat{\text{E}}_{Haj}[Y_i]$ is consistent for $\text{E}[Y_i]$.

Proof: Since ignorability holds, by Theorem 4.2.2,

$$\frac{1}{n}\sum_{i=1}^n \left[\frac{Y_i^* R_i}{P_i}\right] \xrightarrow{p} \text{E}[Y_i].$$

And since

$$\text{E}\left[\frac{R_i}{P_i}\middle|\mathbf{X}_i\right] = \text{E}\left[\frac{R_i}{\Pr(R_i = 1|\mathbf{X}_i)}\middle|\mathbf{X}_i\right] = \frac{\text{E}[R_i|\mathbf{X}_i]}{\Pr(R_i = 1|\mathbf{X}_i)} = \frac{\Pr(R_i = 1|\mathbf{X}_i)}{\Pr(R_i = 1|\mathbf{X}_i)} = 1,$$

by the Law of Iterated Expectations,

$$\text{E}\left[\frac{R_i}{P_i}\right] = \text{E}_{\mathbf{X}_i}\left[\text{E}\left[\frac{R_i}{P_i}\middle|\mathbf{X}_i\right]\right] = \text{E}_{\mathbf{X}_i}[1] = 1.$$

So by the WLLN,

$$\frac{1}{n}\sum_{i=1}^n \frac{R_i}{P_i} \xrightarrow{p} \text{E}\left[\frac{R_i}{P_i}\right] = 1.$$

Thus, by Slutsky's Theorem,

$$\hat{\text{E}}_{Haj}[Y_i] = \frac{\sum_{i=1}^n \frac{Y_i^* R_i}{P_i}}{\sum_{i=1}^n \frac{R_i}{P_i}} = \frac{\frac{1}{n}\sum_{i=1}^n \frac{Y_i^* R_i}{P_i}}{\frac{1}{n}\sum_{i=1}^n \frac{R_i}{P_i}} \xrightarrow{p} \frac{\text{E}[Y_i]}{1} = \text{E}[Y_i]. \ \square$$

In general, IPW estimators tend to be more asymptotically efficient than hot-deck imputation estimators. Note that we still have to estimate the propensity scores. However, IPW estimators have a counterintuitive property: assuming that you have a consistent estimator for the propensity scores, using estimated propensity scores can actually be *more* efficient than using the true propensity scores. The reason for this is that using the estimated probabilities does more to balance the covariates in the sample.

Inference for IPW estimators can be achieved via the bootstrap, re-estimating the propensity scores in each bootstrap sample.

4.2.6 Doubly Robust Estimators

Finally, it is also possible to combine regression with weighting. Estimators of this type are known as *doubly robust (DR)*.[89] Under ignorability, these approaches yield consistent estimates when *either* the regression specification *or* the propensity score specification is correct.

Definition 4.2.3. *The Doubly Robust (DR) Estimator for Missing Data*

Let Y_i and R_i be random variables with $\text{Supp}\left(R_i\right) = \{0, 1\}$, let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(R_i = 1|\mathbf{X}_i)$. Then the doubly robust estimator for $\text{E}\left[Y_i\right]$ is

$$\hat{\text{E}}_{DR}[Y_i] = \frac{1}{n}\sum_{i=1}^{n} \hat{\text{E}}\left[Y_i^*|R_i = 1, \mathbf{X}_i\right] + \frac{1}{n}\sum_{i=1}^{n} \frac{R_i(Y_i^* - \hat{\text{E}}\left[Y_i^*|R_i = 1, \mathbf{X}_i\right])}{\hat{P}_i},$$

where $\hat{\text{E}}\left[Y_i^*|R_i = 1, \mathbf{X}_i\right]$ is an estimator for $\text{E}\left[Y_i^*|R_i = 1, \mathbf{X}_i\right]$ and $\hat{P}_i$ is an estimator for P_i.

This estimator combines a weighting estimator and an imputation estimator. It's really quite brilliant. One simple way to think about the DR estimator is as follows: the first term,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\text{E}}\left[Y_i^*|R_i = 1, \mathbf{X}_i\right],$$

is just the standard regression estimator from Section 4.2.2, while the second term,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R_i(Y_i^* - \hat{\text{E}}\left[Y_i^*|R_i = 1, \mathbf{X}_i\right])}{\hat{P}_i},$$

is an IPW estimator correcting for any "unusual" deviations of the actual data from the imputation estimate. The following theorem states that, under ignorability, the DR estimator is consistent even when one of these components is misspecified.

Theorem 4.2.4. *Consistency of the DR Estimator for Missing Data*

Let Y_i and R_i be random variables with $\text{Supp}\left(R_i\right) = \{0, 1\}$, let $Y_i^* = Y_i R_i + (-99)(1 - R_i)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(R_i = 1|\mathbf{X}_i)$. Then if Y_i is missing at random conditional on $\mathbf{X}_i$, the doubly robust estimator $\hat{\text{E}}_{DR}[Y_i]$ is consistent for $\text{E}\left[Y_i\right]$ if either

[89] See Robins and Rotnitzky (2001) or Kang and Schafer (2007) for more details.

$$\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] \xrightarrow{p} \mathrm{E}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] \quad \textit{or} \quad \hat{P}_i \xrightarrow{p} P_i,$$

Proof: First, suppose that the regression specification is correct, so that

$$\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] \xrightarrow{p} \mathrm{E}\left[Y_i^*|R_i=1,\mathbf{X}_i\right].$$

By ignorability,

$$\mathrm{E}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] = \mathrm{E}\left[Y_i|\mathbf{X}_i\right],$$

so

$$\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] \xrightarrow{p} \mathrm{E}\left[Y_i|\mathbf{X}_i\right].$$

Denote the probability limit of the (possibly misspecified) $\hat{P}_i$ as P_i'. Then by the WLLN,

$$\begin{aligned}
\hat{\mathrm{E}}_{DR}[Y_i] &\xrightarrow{p} \mathrm{E}\left[\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] + \frac{R_i(Y_i^* - \hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right])}{\hat{P}_i}\right] \\
&= \mathrm{E}\left[\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right]\right] + \mathrm{E}\left[\frac{R_i(Y_i^* - \hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right])}{\hat{P}_i}\right] \\
&\xrightarrow{p} \mathrm{E}\left[\mathrm{E}\left[Y_i|\mathbf{X}_i\right]\right] + \mathrm{E}\left[\frac{R_i(Y_i^* - \mathrm{E}\left[Y_i^*|R_i=1,\mathbf{X}_i\right])}{P_i'}\right] \\
&= \mathrm{E}\left[Y_i\right] + \mathrm{E}\left[\frac{R_i(\mathrm{E}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] - \mathrm{E}\left[Y_i^*|R_i=1,\mathbf{X}_i\right])}{P_i'}\right] \\
&= \mathrm{E}\left[Y_i\right] + \mathrm{E}\left[\frac{R_i \cdot 0}{P_i'}\right] \\
&= \mathrm{E}\left[Y_i\right],
\end{aligned}$$

where the third line follows from the CMT and the fourth from the Law of Iterated Expectations.

Now, suppose that the weighting specification is right, so that $\hat{P}_i \xrightarrow{p} P_i$. Denote the probability limit of the (possibly misspecified) $\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right]$ as $\mathrm{E}'[Y_i^*|R_i=1,\mathbf{X}_i]$. We can break apart the second term of the DR estimator:

$$\begin{aligned}
\hat{\mathrm{E}}_{DR}[Y_i] &= \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] + \frac{1}{n}\sum_{i=1}^{n} \frac{R_i(Y_i^* - \hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right])}{\hat{P}_i} \\
&= \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] + \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i^*}{\hat{P}_i} - \frac{1}{n}\sum_{i=1}^{n} \frac{R_i(\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right])}{\hat{P}_i}.
\end{aligned}$$

Then by the WLLN,

$$\begin{aligned}
\hat{\mathrm{E}}_{DR}[Y_i] &\xrightarrow{p} \mathrm{E}\left[\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right] + \frac{R_i Y_i^*}{\hat{P}_i} - \frac{R_i(\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right])}{\hat{P}_i}\right] \\
&= \mathrm{E}\left[\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right]\right] + \mathrm{E}\left[\frac{R_i Y_i^*}{\hat{P}_i}\right] - \mathrm{E}\left[\frac{R_i(\hat{\mathrm{E}}\left[Y_i^*|R_i=1,\mathbf{X}_i\right])}{\hat{P}_i}\right] \\
&\xrightarrow{p} \mathrm{E}\left[\mathrm{E}'[Y_i^*|R_i=1,\mathbf{X}_i]\right] + \mathrm{E}\left[\frac{R_i Y_i^*}{P_i}\right] - \mathrm{E}\left[\frac{R_i(\mathrm{E}'[Y_i^*|R_i=1,\mathbf{X}_i])}{P_i}\right] \\
&= \mathrm{E}\left[\mathrm{E}'[Y_i^*|R_i=1,\mathbf{X}_i]\right] + \mathrm{E}\left[\frac{R_i Y_i^*}{P_i}\right] - \mathrm{E}\left[\frac{R_i}{P_i}\right]\mathrm{E}\left[\mathrm{E}'[Y_i^*|R_i=1,\mathbf{X}_i]\right] \\
&= \mathrm{E}\left[\mathrm{E}'[Y_i^*|R_i=1,\mathbf{X}_i]\right] + \mathrm{E}\left[\frac{R_i Y_i^*}{P_i}\right] - \mathrm{E}\left[\mathrm{E}'[Y_i^*|R_i=1,\mathbf{X}_i]\right] \\
&= \mathrm{E}\left[\frac{R_i Y_i^*}{P_i}\right] \\
&= \mathrm{E}\left[Y_i\right],
\end{aligned}$$

where the third line follows from the CMT. □

In short, if the imputation specification is right, then the probability limit of the second term (the average deviation from the imputation specification conditional on $\mathbf{X}_i$) is zero, leaving us with just the (consistent) imputation estimator. And if weighting specification is right, then the second term is just the standard HT estimator plus a consistent estimator for the negative of the incorrect $\mathrm{E}'[Y_i^*|R_i=1,\mathbf{X}_i]$, which cancels out the first term in the limit and leaves us with just the (consistent) HT estimator.

Inference for the DR estimator can be achieved via the bootstrap, re-estimating the propensity scores and regression fits in each bootstrap sample.

4.2.7 Identification, Estimation, and Assumptions

Remember: all of these estimators rely on the ignorability assumption to produce consistent estimates of $\mathrm{E}\left[Y_i\right]$. No statistical method can make ignorability plausible. The question we have addressed here is how to implement this assumption in a principled manner. All of these methods for implementing the ignorability assumption tend to produce fairly similar estimates—in fact, there exist special cases in which all are exactly equivalent. So don't worry too much about the choice of estimator. Worry about the nonparametric identification conditions (e.g., ignorability). A focus on the estimator is a red herring, and so you should beware of anyone "selling" a method. There's no magic bullet.

5 Causal Inference

Fortunate is one who is able to know the causes of things.

— VIRGIL

What is causal inference? Recall that a feature of a probability distribution is identified if it can be inferred given full knowledge of the distribution of observables. *Causal identification* is just identification of any feature of a distribution that can be interpreted as a causal effect. *Causal inference*, then, is the process of estimating these quantities from data.[90]

The challenges of drawing causal inferences and handling missing data are closely related. Readers will note that the structure and language of this chapter largely mirrors that of Chapter 4. This is no accident, and as we proceed, we will show how the identification results and estimation procedures that we derived for missing data are easily translated to address questions of cause and effect.

5.1 Potential Outcomes

We employ a model of *potential outcomes* (Neyman 1923) known as the Neyman-Rubin Causal Model (NRCM).[91] The NRCM is a powerful and flexible framework for describing claims of cause and effect. There are other formalizations of causality, but the NRCM is

1. the most commonly used framework in political science, biostatistics, and many branches of applied econometrics;
2. completely nonparametric, requiring no distributional assumptions; and
3. in our view, the clearest way to think about causal inference.

Suppose that we want to know the effect of a medication on a patient's self-reported level of pain. The problem is that we cannot observe someone's counterfactual outcome. If the patient takes the medication, we cannot know what she would have reported if she hadn't. If she doesn't take the medication, we cannot know what she would have reported if she had. This is known as the *Fundamental Problem of Causal Inference* (Holland 1986).

5.1.1 Framework

Suppose that we have a binary treatment D_i, so that $D_i = 1$ if unit i receives the treatment and 0 otherwise. Let $Y_i(0)$ denote the potential outcome under control for unit i and $Y_i(1)$ denote the potential outcome

[90] In this book, we have generally used the term "inference" to refer specifically to the quantification of uncertainty. However, the term "causal inference" is commonly used in this broader sense encompassing both estimation and inference, and we will adhere to this established convention.

[91] Also sometimes referred to simply as the Rubin Causal Model.

under treatment for unit i. Then the NRCM implies that

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

where the distribution of (Y_i, D_i) is observed, but the distribution of $(Y_i(0), Y_i(1))$ is not directly observed. Potential outcomes $Y_i(d)$ (for $d \in \{0, 1\}$) can be binary, discrete, continuous, etc., as long as $\text{Supp}\,(Y_i(d)) \subseteq \mathbb{R}$. The population values of $(Y_i(0), Y_i(1), D_i)$ may have any joint distribution (subject to perhaps mild regularity conditions for identification and estimation).

The above equation embeds what is known as the *stable unit treatment value assumption (SUTVA)*. Potential outcomes are assumed to be stable. No matter what, for every unit i, when $D_i = 1$, we observe $Y_i(1)$, and when $D_i = 0$, we observe $Y_i(0)$. This implies that

- there are no unobserved multiple versions of the treatment or control—every unit i has a single outcome value $Y_i(1)$ that is observed when $D_i = 1$ and a single outcome value $Y_i(0)$ that is observed when it $D_i = 0$.
- there is no interference between units—unit i's potential outcomes are not affected by whether or not any other unit j receives the treatment.[92]

The causal effect of treatment for unit i is

$$\tau_i = Y_i(1) - Y_i(0).$$

In general, for reasons that will become apparent later, we will focus on estimating $\text{E}\,[\tau_i]$, the *average treatment effect (ATE)*, also known as the average causal effect (ACE). Note that some texts refer to this quantity as the population average treatment effect (PATE), since we are assuming i.i.d. sampling from some population.[93]

By linearity of expectations, the ATE can be written as

$$\begin{aligned}\text{E}\,[\tau_i] &= \text{E}\,[Y_i(1) - Y_i(0)] \\ &= \text{E}\,[Y_i(1)] - \text{E}\,[Y_i(0)].\end{aligned}$$

This fact is central to causal identification, as we will see shortly.

5.1.2 Ties to Missing Data

We do *not* directly observe i.i.d. draws of $Y_i(0)$ or $Y_i(1)$, and we *never* observe τ_i for any unit i (that's the Fundamental Problem of Causal Inference). We only observe i.i.d. draws of the random vector (Y_i, D_i).

[92] Our recommendation is to not worry too much about SUTVA. There are generalizations of the NRCM that imply that all of our calculations will be meaningful even when SUTVA does not hold, though the interpretation is somewhat more complex.

[93] There exists an alternative formulation of causal inference under the NRCM that does *not* require the assumption of i.i.d. sampling, known as *design-based inference*. See, e.g., Imbens and Rubin (2015). In this book, we focus on causal inference in the i.i.d. sampling context, as it considerably eases exposition.

So the question is: given full knowledge of the distribution of (Y_i, D_i), what assumptions yield what information about the distributions of $Y_i(0)$, $Y_i(1)$, and τ_i?

Causal inference under the NRCM is a missing data problem! (This insight owes to Rubin, which is why we have the 'R' in NRCM.) We only learn about $Y_i(1)$ for treated units and we only learn about $Y_i(0)$ for control units.

Example 5.1.1. *Causal Inference as a Missing Data Problem*

Suppose we had binary outcomes, i.e., $\text{Supp}\left(Y_i(0)\right) \subseteq \{0,1\}$ and $\text{Supp}\left(Y_i(1)\right) \subseteq \{0,1\}$, and therefore $\text{Supp}\left(Y_i\right) \subseteq \{0,1\}$. We might see the following data:

Unit	Y_i	D_i
1	1	1
2	1	0
3	1	1
4	0	1
5	1	1
6	0	0

Given SUTVA, this implies:

Unit	$Y_i(0)$	$Y_i(1)$	D_i
1	?	1	1
2	1	?	0
3	?	1	1
4	?	0	1
5	?	1	1
6	0	?	0

We want to learn about the distributions of $Y_i(0)$ and $Y_i(1)$. But we have a missing data problem. We can identify the joint distribution of (Y_i, D_i) values—with large n, we could learn about the distribution of (Y_i, D_i) to arbitrary precision. But even with large n, we need additional assumptions to learn about the distributions of $Y_i(0)$ and $Y_i(1)$. It's even harder to learn about τ_i.

Recognizing that the Fundamental Problem of Causal Inference is essentially a missing data problem helps us to understand the types of assumptions that we need in order to estimate causal effects. Causal inference under the NRCM is about coming up with ways of "solving" this missing data problem under further assumptions. The approaches that we discuss in this chapter are all completely analogous to those we presented for missing data in Chapter 4.

5.1.3 Bounds

Let's apply Manski bounds to this problem. We only assume that potential outcomes are bounded, i.e., $\forall d \in \{0,1\}, \text{Supp}\left(Y_i(d)\right) \subseteq [a,b]$, for some $a \leq b$. In Example 5.1.1, outcomes are binary, so $\forall d \in \{0,1\}, \text{Supp}\left(Y_i(d)\right) \subseteq \{0,1\} \subseteq [0,1]$.

For the treated units (i.e., those with $D_i = 1$), we observe $Y_i(1)$, but we only know that $0 \leq Y_i(0) \leq 1$. And conversely, for the control units (i.e., those with $D_i = 0$), we observe $Y_i(0)$, but we only know that $0 \leq Y_i(1) \leq 1$.

How can we estimate bounds for $\text{E}\left[\tau_i\right]$? We know that $\text{E}\left[\tau_i\right] = \text{E}\left[Y_i(1)\right] - \text{E}\left[Y_i(0)\right]$, so an upper bound is obtained by estimating an upper bound for $\text{E}\left[Y_i(1)\right]$ and a lower bound for $\text{E}\left[Y_i(0)\right]$. These are computed in the standard way, by plugging in 1 for the missing values of $Y_i(1)$ and 0 for the missing values of $Y_i(0)$:

Unit	$Y_i(0)$	$Y_i(1)$	D_i
1	0	1	1
2	1	1	0
3	0	1	1
4	0	0	1
5	0	1	1
6	0	1	0

Then, using the sample mean as a plug-in estimator for the expected value, our estimate of an upper bound for $\text{E}\left[Y_i(1)\right]$ is $5/6$, and our estimate of a lower bound for $\text{E}\left[Y_i(0)\right]$ is $1/6$. Thus, our estimate of an upper bound for $\text{E}\left[\tau_i\right]$ is $5/6 - 1/6 = 4/6$.

Likewise, to estimate a lower bound for $\text{E}\left[\tau_i\right]$, we estimate a lower bound for $\text{E}\left[Y_i(1)\right]$ and an upper bound for $\text{E}\left[Y_i(0)\right]$:

Unit	$Y_i(0)$	$Y_i(1)$	D_i
1	1	1	1
2	1	0	0
3	1	1	1
4	1	0	1
5	1	1	1
6	0	0	0

Our estimated lower bound for $\text{E}\left[Y_i(1)\right]$ is $3/6$, and our estimated upper bound for $\text{E}\left[Y_i(0)\right]$ is $5/6$. Thus, our estimated lower bound for $\text{E}\left[\tau_i\right]$ is $3/6 - 5/6 = -2/6$.

More generally, we can write down exact sharp bounds for $\text{E}\left[Y_i(0)\right]$, $\text{E}\left[Y_i(1)\right]$, and $\text{E}\left[\tau_i\right]$ in terms of the joint distribution of (Y_i, D_i), which can, in theory, be estimated to arbitrary precision.

Theorem 5.1.1. *Sharp Bounds for the ATE*

Let $Y_i(0)$, $Y_i(1)$, *and* D_i *be random variables such that,* $\forall d \in \{0,1\}, \text{Supp}\left(Y_i(d)\right) \subseteq [a,b]$, *and* $\text{Supp}\left(D_i\right) = \{0,1\}$. *Let* $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$ *and* $\tau_i = Y_i(1) - Y_i(0)$. *Then*

$$\begin{aligned}
\text{E}\left[Y_i(0)\right] \in [&\text{E}\left[Y_i | D_i = 0\right] \Pr(D_i = 0) + a \Pr(D_i = 1), \\
&\text{E}\left[Y_i | D_i = 0\right] \Pr(D_i = 0) + b \Pr(D_i = 1)], \\
\text{E}\left[Y_i(1)\right] \in [&\text{E}\left[Y_i | D_i = 1\right] \Pr(D_i = 1) + a \Pr(D_i = 0), \\
&\text{E}\left[Y_i | D_i = 1\right] \Pr(D_i = 1) + b \Pr(D_i = 0)],
\end{aligned}$$

and

$$\begin{aligned}
\text{E}\left[\tau_i\right] \in [&\text{E}\left[Y_i | D_i = 1\right] \Pr(D_i = 1) + a \Pr(D_i = 0) - (\text{E}\left[Y_i | D_i = 0\right] \Pr(D_i = 0) + b \Pr(D_i = 1)), \\
&\text{E}\left[Y_i | D_i = 1\right] \Pr(D_i = 1) + b \Pr(D_i = 0) - (\text{E}\left[Y_i | D_i = 0\right] \Pr(D_i = 0) + a \Pr(D_i = 1))].
\end{aligned}$$

The proof follows directly from Theorem 4.1.1. Plug-in estimation works as usual and the bootstrap can be used to compute confidence intervals for each bound.

What sort of assumptions do we need to achieve point identification of $\text{E}\left[Y_i(0)\right]$, $\text{E}\left[Y_i(1)\right]$, and $\text{E}\left[\tau_i\right]$? The answer is completely analogous to the case of missing data.

5.1.4 Random Assignment

The strongest nonparametric assumption we can impose is *random assignment* of the treatment. This assumption is analogous to MCAR.

Definition 5.1.1. *Random Assignment*

Let $Y_i(0)$, $Y_i(1)$, *and* D_i *be random variables with* $\text{Supp}\left(D_i\right) = \{0,1\}$. *Let* $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$. *Then* D_i *is randomly assigned if the following conditions hold:*

- $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i$. *(Independence of potential outcomes and treatment)*
- $0 < \Pr(D_i = 1) < 1$. *(Positivity)*

In other words, the distribution of potential outcomes is exactly the same for units that get treated and units that don't. Merely observing whether or not a unit was assigned to treatment reveals nothing about its potential outcomes. Under this assumption, $\text{E}\left[Y_i(0)\right]$, $\text{E}\left[Y_i(1)\right]$, and $\text{E}\left[\tau_i\right]$ are all point identified.

Theorem 5.1.2. *The ATE under Random Assignment*

Let $Y_i(0)$, $Y_i(1)$, *and* D_i *be random variables with* $\text{Supp}\left(D_i\right) = \{0,1\}$. *Let* $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$ *and* $\tau_i = Y_i(1) - Y_i(0)$. *Then if* D_i *is randomly assigned,*

$$\text{E}\left[\tau_i\right] = \text{E}\left[Y_i | D_i = 1\right] - \text{E}\left[Y_i | D_i = 0\right].$$

Proof: By independence (see Theorem 1.4.26),

$$\mathrm{E}\left[Y_i(0)\right] = \mathrm{E}\left[Y_i(0)|D_i = 0\right].$$

And by SUTVA,

$$\mathrm{E}\left[Y_i(0)|D_i = 0\right] = \mathrm{E}\left[Y_i|D_i = 0\right].$$

Thus,

$$\mathrm{E}\left[Y_i(0)\right] = \mathrm{E}\left[Y_i|D_i = 0\right].$$

And by the same logic,

$$\mathrm{E}\left[Y_i(1)\right] = \mathrm{E}\left[Y_i(1)|D_i = 1\right] = \mathrm{E}\left[Y_i|D_i = 1\right].$$

Thus,

$$\begin{aligned}\mathrm{E}\left[\tau_i\right] &= \mathrm{E}\left[Y_i(1) - Y_i(0)\right] \\ &= \mathrm{E}\left[Y_i(1)\right] - \mathrm{E}\left[Y_i(0)\right] \\ &= \mathrm{E}\left[Y_i|D_i = 1\right] - \mathrm{E}\left[Y_i|D_i = 0\right]. \ \square\end{aligned}$$

Under random assignment, the plug-in estimator for $\mathrm{E}\left[Y_i(1)\right]$ is just the sample mean of the treated units:

$$\hat{\mathrm{E}}\left[Y_i(1)\right] = \hat{\mathrm{E}}\left[Y_i|D_i = 1\right] = \frac{\sum_{i=1}^n Y_i D_i}{\sum_{i=1}^n D_i}.$$

Likewise, the plug-in estimator for $\mathrm{E}\left[Y_i(0)\right]$ is the sample mean of the control units:

$$\hat{\mathrm{E}}\left[Y_i(0)\right] = \hat{\mathrm{E}}\left[Y_i|D_i = 0\right] = \frac{\sum_{i=1}^n Y_i(1 - D_i)}{\sum_{i=1}^n (1 - D_i)}.$$

Thus, the plug-in estimator for the ATE is just the difference of these two sample means, which we call the *difference in means estimator*:

$$\hat{\mathrm{E}}_{DM}[\tau_i] = \hat{\mathrm{E}}\left[Y_i(1)\right] - \hat{\mathrm{E}}\left[Y_i(0)\right] = \frac{\sum_{i=1}^n Y_i D_i}{\sum_{i=1}^n D_i} - \frac{\sum_{i=1}^n Y_i(1 - D_i)}{\sum_{i=1}^n (1 - D_i)}.$$

E.g., in Example 5.1.1, our estimate of the ATE would be $3/4 - 1/2 = 1/4$.

Much like MCAR for missing data, we can think of random assignment as allowing us, given large enough n, to impute the unobserved potential outcomes using sample means. Again, consider Example 5.1.1. We have:

Unit	$Y_i(0)$	$Y_i(1)$	D_i
1	?	1	1
2	1	?	0
3	?	1	1
4	?	0	1
5	?	1	1
6	0	?	0

Assuming random assignment, $\mathrm{E}\left[Y_i(1)|D_i = 0\right] = \mathrm{E}\left[Y_i(1)\right]$ and $\mathrm{E}\left[Y_i(0)|D_i = 1\right] = \mathrm{E}\left[Y_i(0)\right]$, which we can estimate using the above plug-in estimators. Thus, we can impute the missing potential outcomes as follows:

Unit	$Y_i(0)$	$Y_i(1)$	D_i
1	$\hat{\mathrm{E}}\left[Y_i(0)\right] = 1/2$	1	1
2	1	$\hat{\mathrm{E}}\left[Y_i(1)\right] = 3/4$	0
3	$\hat{\mathrm{E}}\left[Y_i(0)\right] = 1/2$	1	1
4	$\hat{\mathrm{E}}\left[Y_i(0)\right] = 1/2$	0	1
5	$\hat{\mathrm{E}}\left[Y_i(0)\right] = 1/2$	1	1
6	0	$\hat{\mathrm{E}}\left[Y_i(1)\right] = 3/4$	0

Then the sample means of $Y_i(1)$ and $Y_i(0)$ with the missing values imputed are $3/4$ and $1/2$, respectively, and so our estimate of $\mathrm{E}\left[\tau_i\right]$ is again $3/4 - 1/2 = 1/4$. As with MCAR, under random assignment, imputation is a redundant extra step—we're just using $\hat{\mathrm{E}}\left[Y_i(1)\right]$ and $\hat{\mathrm{E}}\left[Y_i(0)\right]$ to impute the missing values in order to get $\hat{\mathrm{E}}\left[Y_i(1)\right]$ and $\hat{\mathrm{E}}\left[Y_i(0)\right]$ again—but this way of thinking about estimating expected potential outcomes will become useful in the Section 5.1.5.

Even under random assignment, the only meaningful feature of τ_i that we can point identify is $\mathrm{E}\left[\tau_i\right]$. This is because expectations are linear, so we can separate $\mathrm{E}\left[Y_i(1) - Y_i(0)\right]$ into $\mathrm{E}\left[Y_i(1)\right] - \mathrm{E}\left[Y_i(0)\right]$. It is not generally the case that $g(Y_i(1) - Y_i(0)) = g(Y_i(1)) - g(Y_i(0))$. The difference in medians, for instance, is not generally equal to the median difference, so we cannot (nonparametrically) identify the median treatment effect even under random assignment. Likewise for $V(\tau_i)$ and the CDF of τ_i. Although the (marginal) distributions of $Y_i(1)$ and $Y_i(0)$ are identified by random assignment, their *joint* distribution, and thus the distribution of their difference, is not.

Random assignment, or a variant thereof, is what allows us to separate correlation from causation. Random assignment is *not* a given. Consider the following cases, and think about why random assignment would be unlikely to hold:

Unit	Treatment	Outcome
Country	WTO Membership	GNI per Capita
Registered Voter	Voter Turnout Phone Call	Voting
City	Needle Exchange Program	HIV Prevalence

In fact, if you erroneously assume random assignment and take the difference in expected values between treated units and control units, the result can be decomposed as follows:

$$\begin{aligned}
\mathrm{E}\left[Y_i|D_i = 1\right] - \mathrm{E}\left[Y_i|D_i = 0\right] &= \mathrm{E}\left[Y_i(1)|D_i = 1\right] - \mathrm{E}\left[Y_i(0)|D_i = 0\right] \\
&= \mathrm{E}\left[\tau_i + Y_i(0)|D_i = 1\right] - \mathrm{E}\left[Y_i(0)|D_i = 0\right] \\
&= \mathrm{E}\left[\tau_i|D_i = 1\right] + \mathrm{E}\left[Y_i(0)|D_i = 1\right] - \mathrm{E}\left[Y_i(0)|D_i = 0\right],
\end{aligned}$$

where $\mathrm{E}\left[\tau_i|D_i = 1\right]$ is the *average treatment effect among the treated (ATT)*, and $\mathrm{E}\left[Y_i(0)|D_i = 1\right] - \mathrm{E}\left[Y_i(0)|D_i = 0\right]$ is selection bias.

5.1.5 Adding Covariates

If we have additional information on each unit, we can achieve point identification of the ATE under a weaker assumption than random assignment. This assumption is known as *strong ignorability*.

Definition 5.1.2. *Strong Ignorability*

Let $Y_i(0)$, $Y_i(1)$, and D_i be random variables with $\text{Supp}(D_i) = \{0, 1\}$. Let $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$. Then D_i is strongly ignorable conditional on $\mathbf{X}_i$ if the following conditions hold:

- $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | \mathbf{X}_i$. *(Conditional independence assumption)*
- $0 < \Pr(D_i = 1 | \mathbf{X}_i) < 1$. *(Positivity)*

(Note that we are still assuming that treatment is binary. We will relax this assumption in Section 5.1.6.)

Strong ignorability is to random assignment what MAR is to MCAR. It's random assignment *conditional on covariates*. The conditional independence assumption (CIA) states that, among units with the same measured characteristics, the types that receive the treatment and the types that don't are exactly the same in terms of their distribution of potential outcomes.

For example, if we measured age and gender, we could say that, "for people who are 40 years old and male, the CIA implies that there are no systematic differences in potential outcomes between those who get the treatment and those who don't. But we still allow for the possibility that 40-year-olds may be more or less likely to get the treatment than 50-year-olds."

Put another way, the CIA implies that, after accounting for observable background characteristics, knowing whether or not a unit received treatment provides no additional information about a unit's potential outcomes.

Though it is weaker than random assignment, strong ignorability still allows us to point identify $\text{E}[Y_i(0)]$, $\text{E}[Y_i(1)]$, and $\text{E}[\tau_i]$.

Theorem 5.1.3. *The ATE under Strong Ignorability*

Let $Y_i(0)$, $Y_i(1)$, and D_i be random variables with $\text{Supp}(D_i) = \{0, 1\}$. Let $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$ and $\tau_i = Y_i(1) - Y_i(0)$, and let $\mathbf{X}_i$ be a discrete random vector.[94] *Then if D_i is strongly ignorable conditional on $\mathbf{X}_i$,*

$$\text{E}[\tau_i] = \sum_{\mathbf{x}_i} \text{E}[Y_i | D_i = 1, \mathbf{X}_i = \mathbf{x}_i] \Pr(\mathbf{X}_i = \mathbf{x}_i) - \sum_{\mathbf{x}_i} \text{E}[Y_i | D_i = 0, \mathbf{X}_i = \mathbf{x}_i] \Pr(\mathbf{X}_i = \mathbf{x}_i).$$

Proof: By the Law of Iterated Expectations,

$$\text{E}[Y_i(0)] = \text{E}_{\mathbf{X}_i}[\text{E}[Y_i(0) | \mathbf{X}_i]] = \sum_{\mathbf{x}_i} \text{E}[Y_i(0) | \mathbf{X}_i = \mathbf{x}_i] \Pr(\mathbf{X}_i = \mathbf{x}_i).$$

[94] We assume here that $\mathbf{X}_i$ is discrete to ease exposition. This assumption is not necessary, but it lets us work with PMFs instead of PDFs. The continuous case is analogous. Recall that we are writing $\sum_{\mathbf{x}_i}$ as a shorthand for $\sum_{\mathbf{x}_i \in \text{Supp}(\mathbf{X}_i)}$.

And by strong ignorability (and SUTVA),

$$\mathrm{E}\left[Y_i(0)|\mathbf{X}_i\right] = \mathrm{E}\left[Y_i(0)|D_i = 0, \mathbf{X}_i\right] = \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i\right].$$

Thus,

$$\mathrm{E}\left[Y_i(0)\right] = \sum_{\mathbf{x}_i} \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i).$$

And by the same logic,

$$\mathrm{E}\left[Y_i(1)\right] = \sum_{\mathbf{x}_i} \mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i).$$

Thus,

$$\begin{aligned}\mathrm{E}\left[\tau_i\right] &= \mathrm{E}\left[Y_i(1)\right] - \mathrm{E}\left[Y_i(0)\right] \\ &= \sum_{\mathbf{x}_i} \mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i) - \sum_{\mathbf{x}_i} \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i). \ \square\end{aligned}$$

Since we can observe the random vector $(Y_i, D_i, \mathbf{X}_i)$, we can point identify $\mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}_i\right]$, $\mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}_i\right]$, and $\Pr(\mathbf{X}_i = \mathbf{x}_i)$. Thus, under ignorability, $\mathrm{E}\left[\tau_i\right]$ is point identified.

Furthermore, for every $\mathbf{x}_i \in \mathrm{Supp}\left(\mathbf{X}_i\right)$, the *conditional average treatment effect* (CATE) given $\mathbf{X}_i = \mathbf{x}_i$ is point identified.

Theorem 5.1.4. *The CATE under Strong Ignorability*

Let $Y_i(0)$, $Y_i(1)$, and D_i be random variables with $\mathrm{Supp}\left(D_i\right) = \{0, 1\}$. Let $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$ and $\tau_i = Y_i(1) - Y_i(0)$, and let $\mathbf{X}_i$ be a random vector. Then if D_i is strongly ignorable conditional on $\mathbf{X}_i$,

$$\mathrm{E}\left[\tau_i|\mathbf{X}_i\right] = \mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i\right] - \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i\right].$$

Proof: By strong ignorability (and SUTVA),

$$\begin{aligned}&\mathrm{E}\left[Y_i(0)|\mathbf{X}_i\right] = \mathrm{E}\left[Y_i(0)|D_i = 0, \mathbf{X}_i\right] = \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i\right] \text{ and} \\ &\mathrm{E}\left[Y_i(1)|\mathbf{X}_i\right] = \mathrm{E}\left[Y_i(1)|D_i = 1, \mathbf{X}_i\right] = \mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i\right].\end{aligned}$$

Thus, by linearity of conditional expectations,

$$\begin{aligned}\mathrm{E}\left[\tau_i|\mathbf{X}_i\right] &= \mathrm{E}\left[Y_i(1) - Y_i(0)|\mathbf{X}_i\right] \\ &= \mathrm{E}\left[Y_i(1)|\mathbf{X}_i\right] - \mathrm{E}\left[Y_i(0)|\mathbf{X}_i\right] \\ &= \mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i\right] - \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i\right]. \ \square\end{aligned}$$

Note that applying the Law of Iterated Expectations to this result yields the same identification result for the ATE under ignorability as stated in Theorem 5.1.3:

$$\begin{aligned}
\mathrm{E}\left[\tau_i\right] &= \mathrm{E}_{\mathbf{X}_i}\left[\mathrm{E}\left[\tau_i|\mathbf{X}_i\right]\right] \\
&= \sum_{\mathbf{x}_i} \mathrm{E}\left[\tau_i|\mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i) \\
&= \sum_{\mathbf{x}_i} \left(\mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}_i\right] - \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}_i\right]\right) \Pr(\mathbf{X}_i = \mathbf{x}_i) \\
&= \sum_{\mathbf{x}_i} \mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i) - \sum_{\mathbf{x}_i} \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i).
\end{aligned}$$

5.1.6 Generalizing Beyond Binary Treatments

Thus far, we have only discussed causal identification for binary treatments, where each unit either receives the treatment ($D_i = 1$) or doesn't ($D_i = 0$). But suppose we want to consider the causal effects of multivalued or continuous treatments, e.g., the effects of various amounts of cash transfers on hours worked. We can generalize the NRCM to accommodate such cases. Let D_i now be any random variable. Then the generalized potential outcomes model is simply:

$$Y_i = \Big\{ Y_i(d) : D_i = d, \forall d \in \mathrm{Supp}\left(D_i\right).$$

I.e., for every possible value of $d \in \mathrm{Supp}\left(D_i\right)$, each unit i has a corresponding (stable) potential outcome $Y_i(d)$, which is observed when D_i takes on that value. When D_i is binary, this generalization reduces to the original binary potential outcomes model.

Causal effects are then defined as differences in potential outcomes for different treatment levels. For any $d, d' \in \mathrm{Supp}\left(D_i\right)$, the causal effect of moving from d to d' is:

$$\tau_i(d, d') = Y_i(d') - Y_i(d).$$

The key identification assumptions and results are all completely analogous to those we have derived for binary treatments. For example, strong ignorability would entail that

$$\left(Y_i(d)\right)_{d \in \mathrm{Supp}\left(D_i\right)} \perp\!\!\!\perp D_i|\mathbf{X}_i. \text{ (Conditional independence assumption)}$$

We will gloss over the requisite positivity assumption, since it is considerably more opaque in the generalized setting, but it basically ensures the identifiability of all salient expected values. Theorem 5.1.3 then generalizes as follows: under strong ignorability with discrete $\mathbf{X}_i$, for any $d, d' \in \mathrm{Supp}\left(D_i\right)$, the average treatment effect of moving from d to d' is:

$$\mathrm{E}\left[\tau_i(d, d')\right] = \sum_{\mathbf{x}_i} \mathrm{E}\left[Y_i|D_i = d', \mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i) - \sum_{\mathbf{x}_i} \mathrm{E}\left[Y_i|D_i = d, \mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i).$$

The proof is completely analogous to the proof of Theorem 5.1.3. Since we can observe the random vector $(Y_i, D_i, \mathbf{X}_i)$, we can point identify $\mathrm{E}\left[Y_i|D_i = d', \mathbf{X}_i = \mathbf{x}_i\right]$, $\mathrm{E}\left[Y_i|D_i = d, \mathbf{X}_i = \mathbf{x}_i\right]$, and $\Pr(\mathbf{X}_i = \mathbf{x}_i)$.

Thus, under ignorability, for any $d, d' \in \text{Supp}(D_i)$, $\text{E}[\tau_i(d, d')]$ is point identified. Likewise, the generalization of Theorem 5.1.4 is: under strong ignorability, for any $d, d' \in \text{Supp}(D_i)$, the conditional average causal effect given $\mathbf{X}_i$ of moving from d to d' is:

$$\text{E}[\tau_i(d, d')|\mathbf{X}_i] = \text{E}[Y_i|D_i = d', \mathbf{X}_i] - \text{E}[Y_i|D_i = d, \mathbf{X}_i].$$

Again, the proof is completely analogous to the proof of Theorem 5.1.4. Thus, under ignorability, for any $d, d' \in \text{Supp}(D_i)$ and any $\mathbf{x}_i \in \text{Supp}(\mathbf{X}_i)$, $\text{E}[\tau_i(d, d')|\mathbf{X}_i = \mathbf{x}_i]$ is point identified. The Law of Iterated Expectations then straightforwardly implies that for any $d, d' \in \text{Supp}(D_i)$, $\text{E}[\tau_i(d, d')]$ is also point identified under ignorability, as stated above.

5.2 Estimation

We have shown that, under random assignment, estimating $\text{E}[\tau_i]$ is extremely straightforward: the plug-in estimator is just the difference in means estimator. Under strong ignorability, the intuition is essentially the same, but estimation becomes a bit more complicated. In this section, we discuss several ways to implement the strong ignorability assumption to obtain an estimate.

5.2.1 Plug-in Estimation

If we have discrete covariates, then under the strong ignorability assumption, we can consistently estimate $\text{E}[\tau_i]$ with a simple plug-in estimator (rearranged for conciseness):

$$\begin{aligned}\hat{\text{E}}[\tau_i] &= \sum_{\mathbf{x}_i \in \widehat{\text{Supp}}(\mathbf{X}_i)} \left(\hat{\text{E}}[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}_i] - \hat{\text{E}}[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}_i]\right) \widehat{\Pr}(\mathbf{X}_i = \mathbf{x}_i) \\ &= \sum_{\mathbf{x}_i \in \widehat{\text{Supp}}(\mathbf{X}_i)} \left(\frac{\sum_{i=1}^n Y_i D_i \cdot \text{I}(\mathbf{X}_i = \mathbf{x}_i)}{\sum_{i=1}^n D_i \cdot \text{I}(\mathbf{X}_i = \mathbf{x}_i)} - \frac{\sum_{i=1}^n Y_i (1 - D_i) \cdot \text{I}(\mathbf{X}_i = \mathbf{x}_i)}{\sum_{i=1}^n (1 - D_i) \cdot \text{I}(\mathbf{X}_i = \mathbf{x}_i)} \right) \cdot \frac{\sum_{i=1}^n \text{I}(\mathbf{X}_i = \mathbf{x}_i)}{n},\end{aligned}$$

where $\text{I}(\cdot)$ is the indicator function and $\widehat{\text{Supp}}(\mathbf{X}_i)$ is the set of all observed values of $\mathbf{X}_i$ (i.e., the plug-in estimator for the support of $\mathbf{X}_i$).[95] This estimator is simply a weighted average of the difference in means estimators for units with the same covariate values, where the weights are the observed frequencies of those covariate profiles. Note that this requires that we have at least one treated unit and one control unit with $\mathbf{X}_i = \mathbf{x}_i$ for every covariate profile $\mathbf{x}_i \in \widehat{\text{Supp}}(\mathbf{X}_i)$, otherwise we'll have a zero in the denominator.

Consider again Example 5.1.1, but now assume that we also observe a binary covariate X_i for every unit.

[95] This estimator is sometimes known as the *post-stratification* estimator for causal inference.

Unit	$Y_i(0)$	$Y_i(1)$	D_i	X_i
1	?	1	1	1
2	1	?	0	0
3	?	1	1	0
4	?	0	1	1
5	?	1	1	1
6	0	?	0	1

Assuming strong ignorability, the above plug-in estimator yields the estimate

$$\begin{aligned}\hat{\mathrm{E}}\left[\tau_i\right] &= \left(\hat{\mathrm{E}}\left[Y_i | D_i = 1, X_i = 0\right] - \hat{\mathrm{E}}\left[Y_i | D_i = 0, X_i = 0\right]\right) \widehat{\Pr}(X_i = 0) \\ &\quad + \left(\hat{\mathrm{E}}\left[Y_i | D_i = 1, X_i = 1\right] - \hat{\mathrm{E}}\left[Y_i | D_i = 0, X_i = 1\right]\right) \widehat{\Pr}(X_i = 1) \\ &= (1 - 1) \cdot \frac{2}{6} + \left(\frac{2}{3} - 0\right) \cdot \frac{4}{6} \\ &= \frac{4}{9}.\end{aligned}$$

Note that this estimate differs from the random assignment-based estimate of $1/4$.

Like random assignment, we can think of strong ignorability as allowing us to impute the unobserved potential outcomes, only now we impute the missing potential outcome for each unit using the sample mean of units in the other treatment condition *with the same covariate values*. In Example 5.1.1, this yields

Unit	$Y_i(0)$	$Y_i(1)$	D_i	X_i
1	$\hat{\mathrm{E}}\left[Y_i(0) \mid X_i = 1\right] = 0$	1	1	1
2	1	$\hat{\mathrm{E}}\left[Y_i(1) \mid X_i = 0\right] = 1$	0	0
3	$\hat{\mathrm{E}}\left[Y_i(0) \mid X_i = 0\right] = 1$	1	1	0
4	$\hat{\mathrm{E}}\left[Y_i(0) \mid X_i = 1\right] = 0$	0	1	1
5	$\hat{\mathrm{E}}\left[Y_i(0) \mid X_i = 1\right] = 0$	1	1	1
6	0	$\hat{\mathrm{E}}\left[Y_i(1) \mid X_i = 1\right] = 2/3$	0	1

Then the sample means of $Y_i(1)$ and $Y_i(0)$ with the missing values imputed are $7/9$ and $1/3$, respectively, and so our estimate of $\mathrm{E}\left[\tau_i\right]$ is again $7/9 - 1/3 = 4/9$.

Plug-in estimation of this sort works when we have only a small number of possible covariate values. When we have many covariates, or continuous covariates, simple plug-in estimation generally isn't feasible. In the remainder of this chapter, we discuss some other ways of estimating the ATE under strong ignorability.

5.2.2 Regression Estimation

Recall Theorem 5.1.4: under strong ignorability, the CATE given $\mathbf{X}_i$ is point identified as

$$\mathrm{E}\left[\tau_i|\mathbf{X}_i\right] = \mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i\right] - \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i\right].$$

Under strong ignorability, the difference between $\mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i\right]$ and $\mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i\right]$ is a causal effect. But notice: $\mathrm{E}\left[Y_i|D_i, \mathbf{X}_i\right]$ is a conditional expectation function. When strong ignorability holds, some features of the CEF of Y_i are *causal*.

We have an amazing tool for estimating the CEF: OLS regression consistently estimates the best linear predictor of the CEF (where "best" means minimum mean squared error). Suppose we assumed that, at least to a first approximation,

$$\mathrm{E}\left[Y_i|D_i, \mathbf{X}_i\right] = \beta_0 + \beta_1 D_i + \mathbf{X}_i\boldsymbol{\beta}.$$

Then

$$\begin{aligned}\mathrm{E}\left[\tau_i|\mathbf{X}_i\right] &= \mathrm{E}\left[Y_i|D_i = 1, \mathbf{X}_i\right] - \mathrm{E}\left[Y_i|D_i = 0, \mathbf{X}_i\right] \\ &= \beta_0 + \beta_1 \cdot 1 + \mathbf{X}_i\boldsymbol{\beta} - (\beta_0 + \beta_1 \cdot 0 + \mathbf{X}_i\boldsymbol{\beta}) \\ &= \beta_1,\end{aligned}$$

so we can estimate the CATE given $\mathbf{X}_i$ using the OLS coefficient estimate $\hat{\beta}_1$:

$$\hat{\mathrm{E}}\left[\tau_i|\mathbf{X}_i\right] = \hat{\beta}_1.$$

Then by the Law of Iterated Expectations,

$$\mathrm{E}\left[\tau_i\right] = \sum_{\mathbf{x}_i} \mathrm{E}\left[\tau_i|\mathbf{X}_i = \mathbf{x}_i\right] \Pr(\mathbf{X}_i = \mathbf{x}_i) = \sum_{\mathbf{x}_i} \beta_1 \Pr(\mathbf{X}_i = \mathbf{x}_i) = \beta_1 \sum_{\mathbf{x}_i} \Pr(\mathbf{X}_i = \mathbf{x}_i) = \beta_1 \cdot 1 = \beta_1,$$

so we can likewise estimate the ATE with $\hat{\mathrm{E}}\left[\tau_i\right] = \hat{\beta}_1$. Thus, insofar as $\mathrm{E}\left[Y_i|D_i, \mathbf{X}_i\right] \approx \beta_0 + \beta_1 D_i + \mathbf{X}_i\boldsymbol{\beta}$, we can interpret $\hat{\beta}_1$ as a good approximation of the ATE.

Example 5.2.1. *Causal Inference with Regression*

Suppose we had the following data:

Unit	$Y_i(0)$	$Y_i(1)$	D_i	X_{1i}	X_{2i}
1	?	2	1	1	7
2	5	?	0	8	2
3	?	3	1	10	3
4	?	10	1	3	1
5	?	2	1	5	2
6	0	?	0	7	0

We want to fill in all of the missing values with $\hat{\mathrm{E}}\left[Y_i | D_i = d, X_{1i} = x_{1i}, X_{2i} = x_{2i}\right]$. Let's see how we could do this using regression. We'll start by assuming that, at least to a first approximation,

$$\mathrm{E}\left[Y_i | D_i, X_{1i}, X_{2i}\right] = \beta_0 + \beta_1 D_i + \beta_2 X_{1i} + \beta_3 X_{2i}.$$

Then we could estimate the coefficients with OLS and use the resulting equation to impute the missing values.

Unit	$Y_i(0)$	$Y_i(1)$	D_i	X_{1i}	X_{2i}
1	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 \cdot 7$	2	1	1	7
2	5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 8 + \hat{\beta}_3 \cdot 2$	0	8	2
3	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 9 + \hat{\beta}_3 \cdot 3$	3	1	9	3
4	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 3 + \hat{\beta}_3 \cdot 1$	10	1	3	1
5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 5 + \hat{\beta}_3 \cdot 2$	2	1	5	2
6	0	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 7 + \hat{\beta}_3 \cdot 0$	0	7	0

(Remember: these are just expectations for units with the same covariate values; we can't *actually* fill in the missing potential outcomes.) Equivalently, we could instead impute all of the potential outcomes, which will yield the same estimate for $\mathrm{E}\left[\tau_i\right]$ and is more straightforward to implement.

Unit	$Y_i(0)$	$Y_i(1)$	D_i	X_{1i}	X_{2i}
1	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 \cdot 7$	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 \cdot 7$	1	1	7
2	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 8 + \hat{\beta}_3 \cdot 2$	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 8 + \hat{\beta}_3 \cdot 2$	0	8	2
3	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 9 + \hat{\beta}_3 \cdot 3$	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 9 + \hat{\beta}_3 \cdot 3$	1	9	3
4	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 3 + \hat{\beta}_3 \cdot 1$	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 3 + \hat{\beta}_3 \cdot 1$	1	3	1
5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 5 + \hat{\beta}_3 \cdot 2$	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 5 + \hat{\beta}_3 \cdot 2$	1	5	2
6	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 7 + \hat{\beta}_3 \cdot 0$	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 7 + \hat{\beta}_3 \cdot 0$	0	7	0

In either case, the estimate of $\mathrm{E}\left[\tau_i\right]$ is just the difference in means estimator with the imputed values. This estimator would be consistent if ignorability held and if the functional form of the CEF were in fact $\mathrm{E}\left[Y_i | D_i, X_{1i}, X_{2i}\right] = \beta_0 + \beta_1 D_i + \beta_2 X_{1i} + \beta_3 X_{2i}$. In the case without interactions between D_i and the covariates, the resulting estimate of $\mathrm{E}\left[\tau_i\right]$ is just $\hat{\beta}_1$.

When n is small, we likely want to fit a simple approximation so as to reduce the variability of our estimates. You may ask, though, how can we relax the working assumption that $\mathrm{E}\left[Y_i | D_i, X_{1i}, X_{2i}\right] = \beta_0 + \beta_1 D_i + \beta_2 X_{1i} + \beta_3 X_{2i}$? Interactions and polynomials! Recall that, for any continuous CEF, corollaries to the Weierstrass Approximation Theorem (e.g., Theorem 3.4.2) guarantee that we can approximate $\mathrm{E}\left[Y_i | D_i, X_{1i}, X_{2i}\right]$ to arbitrary precision in this way.

For example, we could instead impose the following functional form:

$$\begin{aligned}\mathrm{E}\left[Y_i | D_i, X_{1i}, X_{2i}\right] = \beta_0 &+ \beta_1 D_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{1i}^2 + \beta_5 X_{1i} X_{2i} + \beta_6 X_{2i}^2 \\ &+ \beta_7 D_i X_{1i} + \beta_8 D_i X_{2i} + \beta_9 D_i X_{1i}^2 + \beta_{10} D_i X_{1i} X_{2i} + \beta_{11} D_i X_{2i}^2.\end{aligned}$$

This less restrictive approximation is of course preferable with large n—the worst case scenario is that the extra coefficients are unnecessary. But with small n, we might not want to estimate so many coefficients, as this will result in highly imprecise estimates. We could therefore consider a sieve estimator, where the number of interactions and polynomials grows large as n increases (see Section 3.4.5). When the CEF is continuous over the support of the covariates (a usually minor technical assumption), flexible regression allows us to approximate the CEF arbitrarily well given sufficient data.

There are other ways to generate consistent nonparametric estimators of the CEF. The literature on nonparametric regression is vast. In particular, kernel methods (fitting each observation's associated expectation using only data from close observations) are popular, as are regression trees (choosing "cutpoints" in combinations of the covariates to best predict the outcome). We are not going to cover these, but the logic behind them is not too far from what you might expect. The goal is to flexibly approximate the CEF, and there are many ways to do so.

In general, to quickly produce an estimate of the ATE from any of these nonparametric regression approaches, it is easiest to take predicted values for all potential outcomes and then take the difference in means. The logic of this procedure is clear from the plug-in principle:

$$\mathrm{E}\left[\tau_i\right] = \mathrm{E}\left[Y_i(1)\right] - \mathrm{E}\left[Y_i(0)\right] = \mathrm{E}_{\mathbf{X}_i}\left[\mathrm{E}\left[Y_i | D_i = 1, \mathbf{X}_i\right]\right] - \mathrm{E}_{\mathbf{X}_i}\left[\mathrm{E}\left[Y_i | D_i = 0, \mathbf{X}_i\right]\right],$$

for which a natural plug-in estimator is the difference in means of predicted values:

$$\hat{\mathrm{E}}\left[\tau_i\right] = \hat{\mathrm{E}}\left[Y_i(1)\right] - \hat{\mathrm{E}}\left[Y_i(0)\right] = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i = \mathbf{x}_i\right] - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathrm{E}}\left[Y_i | D_i = 0, \mathbf{X}_i = \mathbf{x}_i\right],$$

where $\hat{\mathrm{E}}\left[Y_i | D_i = d, \mathbf{X}_i = \mathbf{x}_i\right]$ is any consistent estimator of the CEF.

In short, regression works as long as we're willing to make it more flexible as n gets large. In applied practice, it's unlikely that you'll deviate too far from linearity. And as a bonus, when you don't interact D_i with the covariates or if you mean-center all of your covariates, you can simply take the coefficient on D_i as your estimate of the ATE, without needing to take predicted values.

In this section, we have established conditions for using linear regression to obtain principled estimates of causal relationships. This result is at the heart of the difference between our approach to causal inference and the standard treatment in econometrics and statistical modeling. We have established a basis for causal inference with regression *without*:

- making any assumptions about an error distribution (e.g., homoskedasticity, normality, etc.).
- assuming linearity (except as an approximation that can be made arbitrarily flexible).
- discussing "exogeneity," "endogeneity," or "omitted variables bias."

We have invoked a *nonparametric*, *substantive* assumption about the relationship between potential outcomes and treatment. This assumption is strong, but we need not conflate this essential identification assumption with the details of statistical specification. Justifying the conditional independence assumption is the hard part; estimation is comparatively easy.

5.2.3 The Role of The Propensity Score

Before we discuss some other methods of estimating the average treatment effect, we must first establish a fundamental result: "The Central Role of the Propensity Score in Observational Studies for Causal Effects" (Rosenbaum and Rubin 1983). For binary treatments, if strong ignorability holds, then instead of directly conditioning on all of the covariates, it suffices to just condition on a summary measure of treatment and the covariates.

This summary measure is the *propensity score*. In the context of causal inference, the propensity score is the conditional probability of treatment given the covariates. Note that the propensity score for any given covariate profile is simply a function of the covariates: if we know that $\mathbf{X}_i = \mathbf{x}_i$, and we know the full joint distribution of $(D_i, \mathbf{X}_i)$, then we know $\Pr(D_i = 1|\mathbf{X}_i = \mathbf{x}_i)$.

The following theorem states the key properties of the propensity score under strong ignorability.

Theorem 5.2.1. *Strong Ignorability and the Propensity Score*

Let $Y_i(0)$, $Y_i(1)$, and D_i be random variables with $\text{Supp}\,(D_i) = \{0, 1\}$. Let $Y_i = Y_i(1)\cdot D_i + Y_i(0)\cdot(1-D_i)$ and $\tau_i = Y_i(1) - Y_i(0)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(D_i = 1|\mathbf{X}_i)$. Then if D_i is strongly ignorable conditional on $\mathbf{X}_i$,

- $D_i \perp\!\!\!\perp \mathbf{X}_i|P_i$. *(Balance conditional on P_i)*
- $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i|P_i$. *(Conditional independence with respect to P_i)*
- $0 < \Pr(D_i = 1|P_i) < 1$. *(Positivity conditional on P_i)*

Proof: We'll prove the first two properties and leave the third as an exercise to the reader. Note that this proof is identical to that of Theorem 4.2.1, with D_i substituted for R_i and $(Y_i(0), Y_i(1))$ substituted for Y_i.

Balance: Since D_i is binary, its distribution is fully governed by its mean, so

$$D_i \perp\!\!\!\perp \mathbf{X}_i|P_i \quad \Leftrightarrow \quad \text{E}\left[D_i|\mathbf{X}_i, P_i\right] = \text{E}\left[D_i|P_i\right].$$

The propensity score can be written as a function of $\mathbf{X}_i$: $P_i = f(\mathbf{X}_i)$. Thus,

$$\text{E}\left[D_i|\mathbf{X}_i, P_i\right] = \text{E}\left[D_i|\mathbf{X}_i, f(\mathbf{X}_i)\right] = \text{E}\left[D_i|\mathbf{X}_i\right] = P_i.$$

And by the Law of Iterated Expectations,

$$\text{E}\left[D_i|P_i\right] = \text{E}\left[\text{E}\left[D_i|\mathbf{X}_i\right]|P_i\right] = \text{E}\left[P_i|P_i\right] = P_i.$$

Therefore,

$$\text{E}\left[D_i|\mathbf{X}_i, P_i\right] = P_i = \text{E}\left[D_i|P_i\right],$$

and thus $D_i \perp\!\!\!\perp \mathbf{X}_i|P_i$.

Conditional independence: Since D_i is binary,

$$D_i \perp\!\!\!\perp (Y_i(0), Y_i(1))|P_i \quad \Leftrightarrow \quad \mathrm{E}\left[D_i|(Y_i(0), Y_i(1)), P_i\right] = \mathrm{E}\left[D_i|P_i\right].$$

By the Law of Iterated Expectations,

$$\mathrm{E}\left[D_i|(Y_i(0), Y_i(1)), P_i\right] = \mathrm{E}\left[\mathrm{E}\left[D_i|(Y_i(0), Y_i(1)), \mathbf{X}_i\right]|(Y_i(0), Y_i(1)), P_i\right].$$

And by strong ignorability,

$$\mathrm{E}\left[D_i|(Y_i(0), Y_i(1)), \mathbf{X}_i\right] = \mathrm{E}\left[D_i|\mathbf{X}_i\right] = P_i.$$

So by substitution,

$$\mathrm{E}\left[D_i|(Y_i(0), Y_i(1)), P_i\right] = \mathrm{E}\left[P_i|(Y_i(0), Y_i(1)), P_i\right] = P_i = \mathrm{E}\left[D_i|P_i\right],$$

and thus $D_i \perp\!\!\!\perp (Y_i(0), Y_i(1))|P_i$. $\square$

In short, Theorem 5.2.1 says that conditioning on the propensity score is equivalent to conditioning on all of the covariates.

It is important to note that balance on $\mathbf{X}_i$ is a *consequence* of conditioning on the propensity score. I.e., after you have conditioned the propensity score, the conditional distribution of covariates for treated units will be the same as the conditional distribution of covariates for control units. In the context of observational studies, a *balance test* (a statistical test of equality of covariate distributions) merely tests the null hypothesis that we have successfully conditioned on the propensity score. If the distributions of covariates, conditional on the propensity score, for treated units and control units are significantly different, then we have done something wrong. However, balance does *not* imply that that strong ignorability holds. Balance on observable characteristics does not imply balance on unobservable characteristics. Without further assumptions, there exists no general test of strong ignorability.

We have shown that, in order to obtain consistent estimates of the ATE, we do not need to condition on all of the covariates as long as we condition on the propensity score. This result is thought to be useful since it might be difficult to (nonparametrically) condition on a large number of variables (unless we use a simple approximation like OLS). But there is a problem here: with real data, the propensity score *must be estimated*.

How would we estimate the propensity score? By performing some sort of regression of D_i on all of the covariates. We have thus simply shifted the problem from estimating $\mathrm{E}\left[Y_i|D_i, \mathbf{X}_i\right]$ to estimating $\mathrm{E}\left[Y_i|D_i, P_i\right]$ and $P_i = \mathrm{E}\left[D_i|\mathbf{X}_i\right]$. Nevertheless, this is what must be done when working with propensity scores.

5.2.4 Matching

The most common, but certainly not the only, propensity score based method for causal inference is *propensity score matching*, an estimator analogous to hot-deck imputation. Matching is another way

of imputing the missing potential outcomes. Under strong ignorability, matching will typically yield a consistent (but otherwise pretty bad) estimator of the ATE.

There are many types of matching, but we will discuss just one type: one-to-one, nearest-neighbor, with-replacement, propensity score matching. For the moment, we'll ignore the need to estimate the propensity score and just assume that it is known.

Example 5.2.2. *Propensity Score Matching*

Suppose we had the following data:

Unit	$Y_i(0)$	$Y_i(1)$	D_i	P_i
1	?	2	1	0.33
2	5	?	0	0.14
3	?	3	1	0.73
4	?	10	1	0.35
5	?	2	1	0.78
6	0	?	0	0.70

Assume that strong ignorability holds and that our goal is to estimate $\mathrm{E}[\tau_i]$. By Theorem 5.2.1, strong ignorability and SUTVA imply that, $\forall p_i \in \mathrm{Supp}(P_i)$,

$$\mathrm{E}[Y_i(0)|D_i = 1, P_i = p_i] = \mathrm{E}[Y_i(0)|D_i = 0, P_i = p_i] = \mathrm{E}[Y_i|D_i = 0, P_i = p_i] \text{ and}$$
$$\mathrm{E}[Y_i(1)|D_i = 0, P_i = p_i] = \mathrm{E}[Y_i(1)|D_i = 1, P_i = p_i] = \mathrm{E}[Y_i|D_i = 1, P_i = p_i],$$

so we want to impute the missing potential outcomes with estimates of $\mathrm{E}[Y_i|D_i = d, P_i = p_i]$:

Unit	$Y_i(0)$	$Y_i(1)$	D_i	P_i
1	$\hat{\mathrm{E}}[Y_i\|D_i = 0, P_i = 0.33]$	2	1	0.33
2	5	$\hat{\mathrm{E}}[Y_i\|D_i = 1, P_i = 0.14]$	0	0.14
3	$\hat{\mathrm{E}}[Y_i\|D_i = 0, P_i = 0.73]$	3	1	0.73
4	$\hat{\mathrm{E}}[Y_i\|D_i = 0, P_i = 0.35]$	10	1	0.35
5	$\hat{\mathrm{E}}[Y_i\|D_i = 0, P_i = 0.78]$	2	1	0.78
6	0	$\hat{\mathrm{E}}[Y_i\|D_i = 1, P_i = 0.70]$	0	0.70

Then, if our estimator of $\mathrm{E}[Y_i|D_i = d, P_i = p_i]$, is consistent, the difference in means with the missing potential outcomes imputed will be a consistent estimator of the ATE.

There are many imputation-type estimators for causal inference that exploit ignorability. (Regression is itself an imputation estimator.) The procedure for (one-to-one, nearest-neighbor) matching is perhaps the simplest one possible: for each treated unit, find the control unit that is *closest* on P_i and use that unit's outcome to fill in the missing control potential outcome for the treated unit. Likewise, for each control

unit, find the treated unit that is closest on P_i and use that unit's outcome to fill in the missing treatment potential outcome for the control unit.

In this example, this yields:

Unit	$Y_i(0)$	$Y_i(1)$	D_i	P_i
1	5	2	1	0.33
2	5	2	0	0.14
3	0	3	1	0.73
4	5	10	1	0.35
5	0	2	1	0.78
6	0	3	0	0.70

So the matching estimate of $\mathrm{E}[\tau_i]$ is $^{22}/_6 - {}^{15}/_6 = {}^{7}/_6 \approx 1.17$.

All matching estimators are based on some variant of this logic. Some versions match on some other summary metric of $\mathbf{X}_i$. Some do one-to-k matching. Some choose observations to empirically maximize the balance of the matched groups. Some even choose, whenever possible, observations that *exactly* match on all covariates—though of course this requires, at minimum, that $\mathbf{X}_i$ be discrete.

Why is one-to-one, nearest-neighbor matching not such a great idea, even when we assume strong ignorability? Why would you not want to just use the outcome for the closest unit in the other treatment condition to estimate a missing potential outcome? Intuitively, matching only uses a small amount of the information available to impute the missing potential outcomes. In the above example, we used $Y_1(1) = 2$ to impute $Y_2(1)$ because $P_1 = 0.33$ is closest to $P_2 = 0.14$. But $P_4 = 0.35$ is almost just as close to P_1, so $Y_4(1) = 10$ probably also gives us some information about $\mathrm{E}[Y_i(1)|P_i = 0.14]$. So why would we ignore $Y_4(1)$ completely and just use $Y_1(1)$ as our guess for $Y_2(1)$?

The problems with this method get even worse when you have to estimate the propensity score. Suppose you don't know the functional form of $\mathrm{E}[D_i|\mathbf{X}_i = \mathbf{x}_i]$. If you aren't very careful to set up a nonparametrically consistent estimator (e.g., some variant of regression with a growing number of polynomial and interaction terms), then your estimator of P_i may never converge to the right value. Then you won't estimate the ATE consistently, even under strong ignorability. Furthermore, even if you do have a consistent estimator of P_i, you'll still need to account for how the uncertainty of your propensity score estimates contributes to the uncertainty of your estimate of $\mathrm{E}[\tau_i]$. Indeed, matching with estimated propensity scores is not generally root-n consistent—loosely speaking, the bias (and therefore MSE) of the estimator shrinks "too slowly" as n grows large—and the usual methods for constructing confidence intervals, *including the bootstrap*, do not generally work (see, e.g., Abadie and Imbens, 2008).

5.2.5 Weighting Estimators

There are better ways to use the propensity score to estimate $\mathrm{E}[\tau_i]$. *Inverse probability weighted (IPW) estimators* provide one such method. The simplest of these is just another version of the Horvitz-Thompson estimator.

Definition 5.2.1. *The Horvitz-Thompson (HT) Estimator for Causal Inference*

Let $Y_i(0)$, $Y_i(1)$, *and* D_i *be random variables with* $\text{Supp}\,(D_i) = \{0, 1\}$. *Let* $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$ *and* $\tau_i = Y_i(1) - Y_i(0)$, *and let* $\mathbf{X}_i$ *be a random vector. Define the propensity score as* $P_i = \Pr(D_i = 1|\mathbf{X}_i)$, *assuming that all* $0 < P_i < 1$. *Then the Horvitz-Thompson estimator for* $\text{E}\,[\tau_i]$ *is*

$$\hat{\text{E}}_{HT}[\tau_i] = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i D_i}{P_i} - \frac{Y_i(1 - D_i)}{1 - P_i} \right].$$

The Horvitz-Thompson estimator simply weights every treated unit's outcome by the inverse of its probability of being in treatment and every control unit's outcome by the inverse of its probability of being in control and then takes the difference in means (with missing potential outcomes effectively set equal to 0). The following theorem states the key properties of this estimator.

Theorem 5.2.2. *Properties of the HT Estimator for Causal Inference*

Let $Y_i(0)$, $Y_i(1)$, *and* D_i *be random variables with* $\text{Supp}\,(D_i) = \{0, 1\}$. *Let* $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$ *and* $\tau_i = Y_i(1) - Y_i(0)$, *and let* $\mathbf{X}_i$ *be a random vector. Define the propensity score as* $P_i = \Pr(D_i = 1|\mathbf{X}_i)$. *Then if* D_i *is strongly ignorable conditional on* $\mathbf{X}_i$, *the Horvitz-Thompson estimator* $\hat{\text{E}}_{HT}[\tau_i]$ *is unbiased and consistent for* $\text{E}\,[\tau_i]$.

Proof: By definition,

$$\text{E}\left[\frac{Y_i D_i}{P_i} \middle| \mathbf{X}_i \right] = \text{E}\left[Y_i \cdot \frac{D_i}{\Pr(D_i = 1|\mathbf{X}_i)} \middle| \mathbf{X}_i \right].$$

And by SUTVA,

$$\text{E}\left[Y_i \cdot \frac{D_i}{\Pr(D_i = 1|\mathbf{X}_i)} \middle| \mathbf{X}_i \right] = \text{E}\left[Y_i(1) \cdot \frac{D_i}{\Pr(D_i = 1|\mathbf{X}_i)} \middle| \mathbf{X}_i \right].$$

Then since strong ignorability holds, $Y_i(1) \perp\!\!\!\perp D_i|\mathbf{X}_i$, so

$$\begin{aligned}
\text{E}\left[Y_i(1) \cdot \frac{D_i}{\Pr(D_i = 1|\mathbf{X}_i)} \middle| \mathbf{X}_i \right] &= \text{E}\,[Y_i(1)|\mathbf{X}_i] \cdot \text{E}\left[\frac{D_i}{\Pr(D_i = 1|\mathbf{X}_i)} \middle| \mathbf{X}_i \right] \\
&= \text{E}\,[Y_i(1)|\mathbf{X}_i] \cdot \frac{\text{E}\,[D_i|\mathbf{X}_i]}{\Pr(D_i = 1|\mathbf{X}_i)} \\
&= \text{E}\,[Y_i(1)|\mathbf{X}_i] \cdot \frac{\Pr(D_i = 1|\mathbf{X}_i)}{\Pr(D_i = 1|\mathbf{X}_i)} \\
&= \text{E}\,[Y_i(1)|\mathbf{X}_i].
\end{aligned}$$

Thus,

$$\text{E}\left[\frac{Y_i D_i}{P_i} \middle| \mathbf{X}_i \right] = \text{E}\,[Y_i(1)|\mathbf{X}_i].$$

So by the Law of Iterated Expectations,

$$\mathrm{E}\left[Y_i(1)\right] = \mathrm{E}_{\mathbf{X}_i}[\mathrm{E}\left[Y_i(1)|\mathbf{X}_i\right]] = \mathrm{E}_{\mathbf{X}_i}\left[\mathrm{E}\left[\frac{Y_i D_i}{P_i}\middle|\mathbf{X}_i\right]\right] = \mathrm{E}\left[\frac{Y_i D_i}{P_i}\right].$$

And by the same logic,

$$\mathrm{E}\left[Y_i(0)\right] = \mathrm{E}_{\mathbf{X}_i}[\mathrm{E}\left[Y_i(0)|\mathbf{X}_i\right]] = \mathrm{E}_{\mathbf{X}_i}\left[\mathrm{E}\left[\frac{Y_i D_i}{1-P_i}\middle|\mathbf{X}_i\right]\right] = \mathrm{E}\left[\frac{Y_i D_i}{1-P_i}\right].$$

Thus, by linearity of expectations,

$$\mathrm{E}\left[\tau_i\right] = \mathrm{E}\left[Y_i(1) - Y_i(0)\right] = \mathrm{E}\left[Y_i(1)\right] - \mathrm{E}\left[Y_i(0)\right] = \mathrm{E}\left[\frac{Y_i D_i}{P_i}\right] - \mathrm{E}\left[\frac{Y_i D_i}{1-P_i}\right] = \mathrm{E}\left[\frac{Y_i D_i}{P_i} - \frac{Y_i D_i}{1-P_i}\right].$$

This equation suggests a simple plug-in estimator,

$$\hat{\mathrm{E}}_{HT}[\tau_i] = \hat{\mathrm{E}}\left[\frac{Y_i D_i}{P_i} - \frac{Y_i D_i}{1-P_i}\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i D_i}{P_i} - \frac{Y_i(1-D_i)}{1-P_i}\right],$$

which is unbiased and consistent by Theorem 2.1.1 and the WLLN. □

Note that the HT estimator is not necessarily unbiased if P_i is estimated, though it remains consistent so long as the estimator of P_i is consistent.

Intuitively, the logic behind IPW estimators for causal inference is the same as for missing data. Suppose that some units are very unlikely to be assigned to treatment, so they have small propensity scores. These units are likely to be underrepresented in the treatment group and overrepresented in the control group, so we want to up-weight them when they are in treatment and down-weight them when they are in control. The reverse goes for units that are unlikely to be assigned to control: we want to up-weight them when they are in control and down-weight them when they are in treatment. In the case where we have a discrete and finite number of values for $\mathbf{X}_i$, the HT estimator is logically equivalent to the plug-in estimator from Section 5.2.1.

Though the HT estimator is unbiased and consistent, it has high variability in small samples. An alternative that performs better in practice is just another version of the Hajek estimator.

Definition 5.2.2. *The Hajek Estimator for Causal Inference*

Let $Y_i(0)$, $Y_i(1)$, and D_i be random variables with $\mathrm{Supp}\left(D_i\right) = \{0,1\}$. Let $Y_i = Y_i(1)\cdot D_i + Y_i(0)\cdot(1-D_i)$ and $\tau_i = Y_i(1) - Y_i(0)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(D_i = 1|\mathbf{X}_i)$. Then the Hajek estimator for $\mathrm{E}\left[\tau_i\right]$ is

$$\hat{\mathrm{E}}_{Haj}[\tau_i] = \frac{\sum_{i=1}^{n}\frac{Y_i D_i}{P_i}}{\sum_{i=1}^{n}\frac{D_i}{P_i}} - \frac{\sum_{i=1}^{n}\frac{Y_i(1-D_i)}{1-P_i}}{\sum_{i=1}^{n}\frac{1-D_i}{1-P_i}}.$$

With the Hajek estimator, we renormalize the weights to sum to n. This normalization is useful in case we draw an unusually large number of units where P_i is large or small. The Hajek estimator is *not* generally unbiased, but it is consistent and is usually* more efficient than the HT estimator.

Theorem 5.2.3. *Consistency of the Hajek Estimator for Causal Inference*

Let $Y_i(0)$, $Y_i(1)$, and D_i be random variables with $\text{Supp}\,(D_i) = \{0,1\}$. *Let $Y_i = Y_i(1)\cdot D_i + Y_i(0)\cdot(1-D_i)$ and $\tau_i = Y_i(1) - Y_i(0)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(D_i = 1|\mathbf{X}_i)$. Then if D_i is strongly ignorable conditional on $\mathbf{X}_i$, the Hajek estimator $\hat{\text{E}}_{Haj}[\tau_i]$ is consistent for* $\text{E}\,[\tau_i]$.

Proof: Since strong ignorability holds, we know from the proof of Theorem 5.2.2 that

$$\text{E}\,[Y_i(1)] = \text{E}\left[\frac{Y_i D_i}{P_i}\right] \quad \text{and} \quad \text{E}\,[Y_i(0)] = \text{E}\left[\frac{Y_i D_i}{1-P_i}\right].$$

So by the WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i D_i}{P_i} \xrightarrow{p} \text{E}\left[\frac{Y_i D_i}{P_i}\right] = \text{E}\,[Y_i(1)] \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(1-D_i)}{1-P_i} \xrightarrow{p} \text{E}\left[\frac{Y_i D_i}{1-P_i}\right] = \text{E}\,[Y_i(0)].$$

And since

$$\text{E}\left[\frac{D_i}{P_i}\middle|\mathbf{X}_i\right] = \text{E}\left[\frac{D_i}{\Pr(D_i=1|\mathbf{X}_i)}\middle|\mathbf{X}_i\right] = \frac{\text{E}\,[D_i|\mathbf{X}_i]}{\Pr(D_i=1|\mathbf{X}_i)} = \frac{\Pr(D_i=1|\mathbf{X}_i)}{\Pr(D_i=1|\mathbf{X}_i)} = 1,$$

by the Law of Iterated Expectations,

$$\text{E}\left[\frac{D_i}{P_i}\right] = \text{E}_{\mathbf{X}_i}\left[\text{E}\left[\frac{D_i}{P_i}\middle|\mathbf{X}_i\right]\right] = \text{E}_{\mathbf{X}_i}[1] = 1.$$

So by the WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{P_i} \xrightarrow{p} \text{E}\left[\frac{D_i}{P_i}\right] = 1.$$

And by the same logic,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1-D_i}{1-P_i} \xrightarrow{p} \text{E}\left[\frac{1-D_i}{1-P_i}\right] = 1.$$

Thus, by Slutsky's Theorem,

$$\begin{aligned}
\hat{\text{E}}_{Haj}[\tau_i] &= \frac{\sum_{i=1}^{n}\frac{Y_i D_i}{P_i}}{\sum_{i=1}^{n}\frac{D_i}{P_i}} - \frac{\sum_{i=1}^{n}\frac{Y_i(1-D_i)}{1-P_i}}{\sum_{i=1}^{n}\frac{1-D_i}{1-P_i}} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i D_i}{P_i}}{\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{P_i}} - \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(1-D_i)}{1-P_i}}{\frac{1}{n}\sum_{i=1}^{n}\frac{1-D_i}{1-P_i}} \\
&\xrightarrow{p} \frac{\text{E}\,[Y_i(1)]}{1} - \frac{\text{E}\,[Y_i(0)]}{1} \\
&= \text{E}\,[Y_i(1)] - \text{E}\,[Y_i(0)] \\
&= \text{E}\,[Y_i(1) - Y_i(0)] = \text{E}\,[\tau_i]. \ \square
\end{aligned}$$

In general, IPW estimators tend to be more asymptotically efficient than matching estimators. Note that we still have to estimate the propensity scores. However, IPW estimators have a counterintuitive property: assuming that you have a consistent estimator for the propensity scores, using estimated propensity scores can actually be more efficient than using the true propensity scores. The reason for this is that using the estimated probabilities does more to balance the covariates in the sample.

Inference for IPW estimators can be achieved via the bootstrap, re-estimating the propensity scores in each bootstrap sample.

5.2.6 Doubly Robust Estimators

Finally, it is also possible to combine regression with weighting specifications using a *doubly robust (DR)* estimator. Under strong ignorability, these approaches yield consistent estimates when *either* the regression specification *or* the propensity score specification is correct.

Definition 5.2.3. *The Doubly Robust (DR) Estimator for Causal Inference*

Let $Y_i(0)$, $Y_i(1)$, *and* D_i *be random variables with* $\text{Supp}\,(D_i) = \{0, 1\}$. *Let* $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$ *and* $\tau_i = Y_i(1) - Y_i(0)$, *and let* $\mathbf{X}_i$ *be a random vector. Define the propensity score as* $P_i = \Pr(D_i = 1|\mathbf{X}_i)$. *Then the doubly robust estimators for* $\text{E}\,[Y_i(1)]$, $\text{E}\,[Y_i(0)]$, *and* $\text{E}\,[\tau_i]$, *respectively, are*

- $\hat{\text{E}}_{DR}[Y_i(1)] = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \hat{\text{E}}\,[Y_i|D_i = 1, \mathbf{X}_i] + \dfrac{1}{n}\sum_{i=1}^{n} \dfrac{D_i(Y_i - \hat{\text{E}}\,[Y_i|D_i = 1, \mathbf{X}_i])}{\hat{P}_i}$,
- $\hat{\text{E}}_{DR}[Y_i(0)] = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \hat{\text{E}}\,[Y_i|D_i = 0, \mathbf{X}_i] + \dfrac{1}{n}\sum_{i=1}^{n} \dfrac{(1 - D_i)(Y_i - \hat{\text{E}}\,[Y_i|D_i = 0, \mathbf{X}_i])}{1 - \hat{P}_i}$,
- $\hat{\text{E}}_{DR}[\tau_i] = \hat{\text{E}}_{DR}[Y_i(1)] - \hat{\text{E}}_{DR}[Y_i(0)]$,

where, $\forall d \in \{0, 1\}$, $\hat{\text{E}}\,[Y_i|D_i = d, \mathbf{X}_i]$ *is an estimator for* $\text{E}\,[Y_i|D_i = d, \mathbf{X}_i]$, *and* $\hat{P}_i$ *is an estimator for* P_i.

As before, this estimator combines a weighting estimator and an imputation estimator. One simple way to think about the DR estimator is as follows: the first term of $\hat{\text{E}}_{DR}[Y_i(1)]$,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\text{E}}\,[Y_i|D_i = 1, \mathbf{X}_i],$$

is just the standard regression estimator from Section 5.2.2, while the second term,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{D_i(Y_i - \hat{\text{E}}\,[Y_i|D_i = 1, \mathbf{X}_i])}{\hat{P}_i},$$

is an IPW estimator correcting for any "unusual" deviations of the actual data from the imputation estimate. The same is true of $\hat{\text{E}}_{DR}[Y_i(0)]$. The following theorem states that, under strong ignorability, the DR estimator is consistent even when one of these components is misspecified.

Theorem 5.2.4. *Consistency of the DR Estimator for Causal Inference*

Let $Y_i(0)$, $Y_i(1)$, and D_i be random variables with $\text{Supp}\,(D_i) = \{0,1\}$. *Let $Y_i = Y_i(1)\cdot D_i + Y_i(0)\cdot(1-D_i)$ and $\tau_i = Y_i(1) - Y_i(0)$, and let $\mathbf{X}_i$ be a random vector. Define the propensity score as $P_i = \Pr(D_i = 1|\mathbf{X}_i)$. Then if D_i is strongly ignorable conditional on $\mathbf{X}_i$, the doubly robust estimator $\hat{\text{E}}_{DR}[\tau_i]$ is consistent for* $\text{E}\,[\tau_i]$ *if either*

$$\forall d \in \{0,1\}, \hat{\text{E}}\,[Y_i|D_i = d, \mathbf{X}_i] \xrightarrow{p} \text{E}\,[Y_i|D_i = d, \mathbf{X}_i] \quad \textit{or} \quad \hat{P}_i \xrightarrow{p} P_i.$$

Proof: First, suppose that the regression specification is correct, so that, $\forall d \in \{0,1\}$,

$$\hat{\text{E}}\,[Y_i|D_i = d, \mathbf{X}_i] \xrightarrow{p} \text{E}\,[Y_i|D_i = d, \mathbf{X}_i].$$

By ignorability,

$$\text{E}\,[Y_i|D_i = 1, \mathbf{X}_i] = \text{E}\,[Y_i(1)|\mathbf{X}_i],$$

so

$$\hat{\text{E}}\,[Y_i|D_i = 1, \mathbf{X}_i] \xrightarrow{p} \text{E}\,[Y_i(1)|\mathbf{X}_i].$$

Denote the probability limit of the (possibly misspecified) $\hat{P}_i$ as P_i'. Then by the WLLN,

$$\begin{aligned}
\hat{\text{E}}_{DR}[Y_i(1)] &\xrightarrow{p} \text{E}\left[\hat{\text{E}}\,[Y_i|D_i = 1, \mathbf{X}_i] + \frac{D_i(Y_i - \hat{\text{E}}\,[Y_i|D_i = 1, \mathbf{X}_i])}{\hat{P}_i}\right] \\
&= \text{E}\left[\hat{\text{E}}\,[Y_i|D_i = 1, \mathbf{X}_i]\right] + \text{E}\left[\frac{D_i(Y_i - \hat{\text{E}}\,[Y_i|D_i = 1, \mathbf{X}_i])}{\hat{P}_i}\right] \\
&\xrightarrow{p} \text{E}\left[\text{E}\,[Y_i(1)|\mathbf{X}_i]\right] + \text{E}\left[\frac{D_i(Y_i - \text{E}\,[Y_i|D_i = 1, \mathbf{X}_i])}{P_i'}\right] \\
&= \text{E}\,[Y_i(1)] + \text{E}\left[\frac{D_i(\text{E}\,[Y_i|D_i = 1, \mathbf{X}_i] - \text{E}\,[Y_i|D_i = 1, \mathbf{X}_i])}{P_i'}\right] \\
&= \text{E}\,[Y_i(1)] + \text{E}\left[\frac{D_i \cdot 0}{P_i'}\right] \\
&= \text{E}\,[Y_i(1)],
\end{aligned}$$

where the third line follows from the CMT and the fourth from the Law of Iterated Expectations.

And by the same logic,

$$\hat{\text{E}}_{DR}[Y_i(0)] \xrightarrow{p} \text{E}\,[Y_i(0)].$$

Thus, by Slutsky's Theorem,

$$\hat{\text{E}}_{DR}[\tau_i] = \hat{\text{E}}_{DR}[Y_i(1)] - \hat{\text{E}}_{DR}[Y_i(0)] \xrightarrow{p} \text{E}\,[Y_i(1)] - \text{E}\,[Y_i(0)] = \text{E}\,[Y_i(1) - Y_i(0)] = \text{E}\,[\tau_i].$$

Now, suppose that the weighting specification is right, so that $\hat{P}_i \xrightarrow{p} P_i$. Denote the probability limit of the (possibly misspecified) $\hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i\right]$ as $\mathrm{E}'[Y_i | D_i = 1, \mathbf{X}_i]$. We can break apart the second term of the DR estimator for $\mathrm{E}\left[Y_i(1)\right]$:

$$\begin{aligned}
\hat{\mathrm{E}}_{DR}[Y_i(1)] &= \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i\right] + \frac{1}{n}\sum_{i=1}^{n} \frac{D_i(Y_i - \hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i\right])}{\hat{P}_i} \\
&= \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i\right] + \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{\hat{P}_i} - \frac{1}{n}\sum_{i=1}^{n} \frac{D_i(\hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i\right])}{\hat{P}_i}.
\end{aligned}$$

Then by the WLLN,

$$\begin{aligned}
\hat{\mathrm{E}}_{DR}[Y_i(1)] &\xrightarrow{p} \mathrm{E}\left[\hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i\right] + \frac{D_i Y_i}{\hat{P}_i} - \frac{D_i(\hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i\right])}{\hat{P}_i}\right] \\
&= \mathrm{E}\left[\hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i\right]\right] + \mathrm{E}\left[\frac{D_i Y_i}{\hat{P}_i}\right] - \mathrm{E}\left[\frac{D_i(\hat{\mathrm{E}}\left[Y_i | D_i = 1, \mathbf{X}_i\right])}{\hat{P}_i}\right] \\
&\xrightarrow{p} \mathrm{E}\left[\mathrm{E}'[Y_i | D_i = 1, \mathbf{X}_i]\right] + \mathrm{E}\left[\frac{D_i Y_i}{P_i}\right] - \mathrm{E}\left[\frac{D_i(\mathrm{E}'[Y_i | D_i = 1, \mathbf{X}_i])}{P_i}\right] \\
&= \mathrm{E}\left[\mathrm{E}'[Y_i | D_i = 1, \mathbf{X}_i]\right] + \mathrm{E}\left[\frac{D_i Y_i}{P_i}\right] - \mathrm{E}\left[\frac{D_i}{P_i}\right]\mathrm{E}\left[\mathrm{E}'[Y_i | D_i = 1, \mathbf{X}_i]\right] \\
&= \mathrm{E}\left[\mathrm{E}'[Y_i | D_i = 1, \mathbf{X}_i]\right] + \mathrm{E}\left[\frac{D_i Y_i}{P_i}\right] - \mathrm{E}\left[\mathrm{E}'[Y_i | D_i = 1, \mathbf{X}_i]\right] \\
&= \mathrm{E}\left[\frac{D_i Y_i}{P_i}\right] \\
&= \mathrm{E}\left[Y_i(1)\right],
\end{aligned}$$

where the third line follows from the CMT.

And by the same logic,

$$\hat{\mathrm{E}}_{DR}[Y_i(0)] \xrightarrow{p} \mathrm{E}\left[Y_i(0)\right].$$

Thus, by Slutsky's Theorem,

$$\hat{\mathrm{E}}_{DR}[\tau_i] = \hat{\mathrm{E}}_{DR}[Y_i(1)] - \hat{\mathrm{E}}_{DR}[Y_i(0)] \xrightarrow{p} \mathrm{E}\left[Y_i(1)\right] - \mathrm{E}\left[Y_i(0)\right] = \mathrm{E}\left[Y_i(1) - Y_i(0)\right] = \mathrm{E}\left[\tau_i\right]. \ \square$$

Inference for the DR estimator can be achieved via the bootstrap, re-estimating the propensity scores and regression fits in each bootstrap sample.

5.2.7 Identification, Estimation, and Assumptions

The same advice from Section 4.2.7 applies here: all of these estimators rely on the strong ignorability assumption to produce consistent estimates of $\mathrm{E}[\tau_i]$. No statistical method can make strong ignorability plausible. The question we have addressed here is how to implement this assumption in a principled manner. All of these methods for implementing the strong ignorability assumption tend to produce fairly similar estimates—in fact, there exist special cases in which all are exactly equivalent. So don't worry so much about the choice of estimator. Worry about the nonparametric identification conditions (e.g., ignorability). A focus on the estimator is a red herring, and so you should beware of anyone "selling" a method. There's no magic bullet.

Given a well-identified study, OLS is probably good enough. It's clear, it's hard to screw up, and it gives you sensible estimates. This is what we recommend. Alternatively, if you want or need to employ a more "high-tech" estimation approach, use a DR estimator with reasonably flexible specifications for the imputation and propensity score estimates. These are arguably the best estimators around, and you would be hard pressed to do better—but they don't solve the identification problem.

Estimation is easy. Identification is hard. If you can't argue for strong ignorability convincingly, then your causal effect estimate is not credible. No amount of fancy-pants statistical razzle-dazzle can get around this. But if you know how the treatment is allocated, then you can identify causal effects. You know that, for at least some classes of people, who received and who did not receive the treatment is a matter of random chance.

How, then, should you approach causal inference in observational research?

1. Think hard about whether or not you have well-defined counterfactuals, and whether or not you're really measuring the "treatment" that you think you are measuring. (E.g., what does "the causal effect of democracy" even mean?)

2. Understand processes well. If you know why some units get the treatment and some don't, this will provide you with a basis for inference. This may mean *reading books*, *conducting interviews*, and *thinking hard*. Then you may be able to invoke strong ignorability and enumerate its implications somewhat convincingly.

3. Write down the identification conditions. If you can't write down an identification result, you probably don't have a valid research design for causal inference. When you actually perform an observational study, write down the target quantities in terms of the CEF. As a first approximation, use OLS to estimate the CEF and get a sense of what's in the data.

5.2.8 Balance Testing

You may ask: are there any ways to *test* the causal identification assumptions underlying an observational study? The answer, technically, is no. The key assumption for causal inference is always some version of the conditional independence assumption. Because of the Fundamental Problem of Causal Inference, we

can *never* observe whether the joint distribution of potential outcomes for treated units is different from that of control units (conditional on covariates).

However, depending on the substance of the study in question, it is sometimes possible to present statistical evidence that may strengthen the credibility of the research design. Suppose that we have some additional covariates $\mathbf{W}_i$ that we don't think we need to condition on and that are not causally affected by the treatment assignment.

Then we can test the null hypothesis that

$$\mathbf{W}_i \perp\!\!\!\perp D_i | \mathbf{X}_i.$$

This is known as a *balance test*. One simple way to implement this is by performing an F-test. An F-test compares a regression of D_i on $\mathbf{W}_i$ and $\mathbf{X}_i$ with a regression of D_i on $\mathbf{X}_i$ alone, and asks: does adding $\mathbf{W}_i$ to the regression explain more of the remaining variation in D_i than we would expect by chance? If so, then we can reject the null hypothesis that $\mathbf{W}_i$ is independent of D_i conditional on $\mathbf{X}_i$.

Substantively, we may not be willing to believe that $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | \mathbf{X}_i$ holds if it is statistically unlikely that $\mathbf{W}_i \perp\!\!\!\perp D_i | \mathbf{X}_i$ holds. I.e., if we can show that, even conditional on $\mathbf{X}_i$, treated units are systematically different from control units in at least some ways, then we should worry that they might be systematically different in terms of their potential outcomes as well.

Balance tests of this sort should not be confused with balance tests after matching or inverse probability weighting. These are used to assess whether or not you have successfully implemented the matching or weighting algorithm. By Theorem 5.2.1, if you have the right propensity score specification, you know that the "balancing property" ensures that you will have statistical independence of D_i and $\mathbf{X}_i$.

5.3 More on Causal Inference with Regression

In Section 5.2.2, we argued for the use of regression to estimate the ATE under strong ignorability as follows:

- The CIA implies a causal interpretation to the CEF.
- The CEF is "probably" close to linear in D_i and $\mathbf{X}_i$.
- Thus, the OLS coefficient on D_i is "probably" a decent approximation of the ATE.

There is, however, an alternative way to think about OLS as an estimator of causal effects, one that helps to clarify several important issues we have yet to discuss. Recall from the FWL Theorem that, for the regression specification $\mathrm{E}\left[Y_i | D_i, \mathbf{X}_i\right] = \beta_0 + \beta_1 D_i + \mathbf{X}_i \boldsymbol{\beta}$,

$$\hat{\beta}_1 = \frac{\widehat{\mathrm{Cov}}\left(Y_i^r, D_i^r\right)}{\hat{V}(D_i^r)}.$$

In other words, the uninteracted regression estimate of the ATE identifies off of variation in the treatment that is not (linearly) explained by the covariates. This construction will help us to understand the operating characteristics of regression for causal inference under different sets of assumptions and specifications.

5.3.1 Covariance Adjustment

Recall from Section 3.2.3 that, for the uninteracted regression $\hat{\mathrm{E}}[Y_i|D_i, \mathbf{X}_i] = \hat{\beta}_0 + \hat{\beta}_1 D_i + \mathbf{X}_i\hat{\boldsymbol{\beta}}$, given sufficiently large n,

$$V(\hat{\beta}_1) \approx \frac{1}{n}\frac{V(e_i D_i^r)}{V(D_i^r)^2}.$$

If we are concerned with asymptotic efficiency in estimating $\hat{\beta}_1$, this asymptotic approximation tells us something about how we want to specify our regression: in general, we want to keep the variance of D_i^r *high* and the variance of e_i *low*. I.e., we want to explain the *outcome* well, but not at the cost of predicting the *treatment* too well. Of course, we also want to be sure to that our regression is consistent.

Suppose we had a randomly assigned binary treatment D_i and a set of pre-treatment covariates $\mathbf{X}_i$, so that

$$(Y_i(0), Y_i(1), \mathbf{X}_i) \perp\!\!\!\perp D_i.$$

Now, we know that if we estimated $\mathrm{E}[Y_i|D_i] = \beta_0 + \beta_1 D_i$ using OLS, $\hat{\beta}_1$ would be consistent for $\mathrm{E}[\tau_i]$. However, we can use $\mathbf{X}_i$ to improve the efficiency of our estimates. Suppose that we instead estimated $\mathrm{E}[Y_i|D_i, \mathbf{X}_i] = \alpha_0 + \alpha_1 D_i + \mathbf{X}_i\hat{\boldsymbol{\alpha}}$. Then our estimate $\hat{\alpha}_1$ is consistent for $\mathrm{E}[\tau_i]$, *and* it is usually* the case that, asymptotically, $V(\hat{\alpha}_1) \leq V(\hat{\beta}_1)$. (There are some odd conditions under which including $\mathbf{X}_i$ under random assignment of D_i can decrease precision. These are rare in practice.)

Why does this work? Since $\mathbf{X}_i \perp\!\!\!\perp D_i$, $\mathbf{X}_i$ does not explain any of the variation in D_i, so in large samples, partialling out $\mathbf{X}_i$ leaves us with $V(D_i^r)$ that's no smaller than $V(D_i)$. But if we residualize off $\mathbf{X}_i$, it can lower the variance of e_i relative to the regression without $\mathbf{X}_i$, thus improving the precision of our estimate.

Thus, even if you have a randomly assigned treatment, i.e. $(Y_i(0), Y_i(1), \mathbf{X}_i) \perp\!\!\!\perp D_i$, you probably still want to control for $\mathbf{X}_i$ when estimating the ATE with regression—if not in your main analysis, then at least as a secondary specification. This is called *covariance adjustment*. It usually* increases precision, and is therefore generally a good idea.

5.3.2 Bad Controls

Adjusting for covariates doesn't always help, though. For example, it is possible that

- $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i$, but
- $(Y_i(0), Y_i(1)) \not\perp\!\!\!\perp D_i|X_i$.

When this happens, the regression estimator $\hat{\mathrm{E}}[Y_i|D_i] = \hat{\beta}_0 + \hat{\beta}_1 D_i$ yields a consistent estimator of the ATE, but $\hat{\mathrm{E}}[Y_i|D_i, X_i] = \hat{\alpha}_0 + \hat{\alpha}_1 D_i + \hat{\alpha}_2 X_i$ does not. The classic case of this is when X_i is *post-treatment*. For example, if D_i has a *causal effect* on X_i, then controlling for X_i can cause inconsistency, as the following example illustrates.

Example 5.3.1. *Bad Controls*

Suppose that

$$f(Y_i(0), Y_i(1), X_i(1), X_i(0), D_i) = \begin{cases} 1/6 & : Y_i(0) = 0, Y_i(1) = 0, X_i(1) = 1, X_i(0) = 0, D_i = 0 \\ 1/6 & : Y_i(0) = 1, Y_i(1) = 1, X_i(1) = 0, X_i(0) = 1, D_i = 0 \\ 1/6 & : Y_i(0) = 1, Y_i(1) = 1, X_i(1) = 1, X_i(0) = 1, D_i = 0 \\ 1/6 & : Y_i(0) = 0, Y_i(1) = 0, X_i(1) = 1, X_i(0) = 0, D_i = 1 \\ 1/6 & : Y_i(0) = 1, Y_i(1) = 1, X_i(1) = 0, X_i(0) = 1, D_i = 1 \\ 1/6 & : Y_i(0) = 1, Y_i(1) = 1, X_i(1) = 1, X_i(0) = 1, D_i = 1 \\ 0 & : otherwise \end{cases}$$

where $X_i = X_i(1)D_i + X_i(0)(1 - D_i)$. In this case, we have

- random assignment of D_i: $(Y_i(0), Y_i(1), X_i(1), X_i(0)) \perp\!\!\!\perp D_i$;
- no effect of D_i on Y_i whatsoever: $Y_i(1) = Y_i(0)$ always; and
- an effect of D_i on X_i.

Then the distribution of observables is:

$$f(Y_i, X_i, D_i) = \begin{cases} 1/6 & : Y_i = 0, X_i = 0, D_i = 0 \\ 2/6 & : Y_i = 1, X_i = 1, D_i = 0 \\ 1/6 & : Y_i = 0, X_i = 1, D_i = 1 \\ 1/6 & : Y_i = 1, X_i = 0, D_i = 1 \\ 1/6 & : Y_i = 1, X_i = 1, D_i = 1 \\ 0 & : otherwise \end{cases}$$

The regression that does not control for X_i is just fine: β_1 from $\mathrm{E}[Y_i|D_i] = \beta_0 + \beta_1 D_i$ is 0. But α_1 from $\mathrm{E}[Y_i|D_i, X_i] = \alpha_0 + \alpha_1 D_i + \alpha_2 X_i$ is 0.25. The regression estimate is now asymptotically biased because we have controlled for X_i. Additional controls can introduce inconsistency, particularly when they are post-treatment.

Never condition on post-treatment variables (unless you have a very good reason to, which you usually* do not). It is a ***very*** *bad idea*, and can even undo a randomly assigned treatment. Note that "mediation analysis" is logically equivalent to conditioning on a post-treatment variable. This is also a ***very*** *bad idea*.

5.3.3 Collinearity

Even when controlling for additional covariates does not introduce inconsistency, it can still increase the standard error of our estimator. For example, suppose the following conditions were to hold:

- $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i$ (random assignment of the treatment),
- $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | X_i$ (conditional independence given X_i),
- $(Y_i(0), Y_i(1)) \perp\!\!\!\perp X_i$ (after accounting for the treatment, X_i does not predict outcomes), and
- $D_i \not\perp\!\!\!\perp X_i$ (X_i predicts D_i).

In this case, since random assignment holds, we don't *need* to control for X_i to get a consistent estimate of the ATE. And since the CIA holds, we will still get a consistent estimate of the ATE if we *do* control for X_i. However, suppose further that $|\rho(X_i, D_i)|$ is close to 1. What happens if we control for X_i?

$$V(\hat{\beta}_1) \approx \frac{1}{n} \frac{V(e_i D_i^r)}{V(D_i^r)^2}.$$

Since X_i is highly correlated with D_i, the variance of D_i^r (i.e. the variation in D_i not explained by X_i) is very small. But since X_i does not predict outcomes after accounting for D_i, the variance of e_i remains relatively large. Thus, the denominator is very small relative to the numerator, so we have a huge standard error for $\hat{\beta}_1$. This is an instance of collinearity. As we saw in Section 3.2.4, imprecision due to collinearity is fundamentally just a problem of having too few observations, or "micronumerosity."

How would we know that we don't need to condition on X_i, especially given that X_i predicts D_i? This is ultimately a substantive question. If we can justify the CIA on substantive grounds, but we do not have good reason to believe that (unconditional) random assignment holds, then we are left with a problem of micronumerosity. Thus, in such a circumstance, we would typically still recommend conditioning on X_i, and incurring the cost of a larger variance.

5.3.4 Regression Weights

We've focused on the case where the CEF is indeed approximately linear. But what does regression give us when there are interactions between D_i and $\mathbf{X}_i$ that we haven't accounted for? The short answer is that we get a reweighted average causal effect, where the weights are proportional to the amount of variation in D_i that is not explained by $\mathbf{X}_i$. The following theorem states this result formally for binary treatments.

Theorem 5.3.1. *Uninteracted Regression is Consistent for a Reweighted ATE*

Let $Y_i(0)$, $Y_i(1)$, and D_i be random variables with $\text{Supp}(D_i) = \{0, 1\}$. Let $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$ and $\tau_i = Y_i(1) - Y_i(0)$, and let $\mathbf{X}_i$ be a random vector. Then if D_i is strongly ignorable conditional on $\mathbf{X}_i$, the regression estimator $\hat{\text{E}}[Y_i | D_i, \mathbf{X}_i] = \hat{\beta}_0 + \hat{\beta}_1 D_i + \mathbf{X}_i \hat{\boldsymbol{\beta}}$ satisfies

$$\hat{\beta}_1 \xrightarrow{p} \frac{\text{E}[w_i \tau_i]}{\text{E}[w_i]},$$

where $w_i = D_i^{r\,2} = (D_i - \mathrm{E}\left[D_i|\mathbf{X}_i\right])^2$.

A proof of a more general version of this theorem (allowing for non-binary treatments) is given by Aronow and Samii (in press), but an intuition follows from applying the FWL Theorem and examining the sources of unexplained variability in D_i. See Angrist and Krueger (1999) or Angrist and Pischke (2009, Ch. 3) for more details.

5.4 Extensions

We conclude by noting that our treatment of causal inference is incomplete with respect to the myriad ways that point identification of causal parameters can be achieved. At their core, however, most such approaches for causal inference rely on an assumption resembling ignorability. We consider here some examples: Instrumental variables strategies typically require an ignorable *instrument*, a variable that affects the outcome only through the channel of one particular treatment. Regression discontinuity designs typically require an assumption closely resembling ignorability at a single point in some *forcing variable* that determines whether or not a unit is treated. Approaches that attempt to use panel (longitudinal) data to achieve point identification usually depend on some type of ignorability assumption, though there are many that can be invoked: ignorability conditional on unit (*fixed effects* estimation), ignorability conditional on behavior and treatments in prior periods (*sequential ignorability*), or an assumption that treatment is ignorable with respect to trends in potential outcomes (*difference-in-difference* estimation).[96]

The key ideas expressed in this chapter are readily extendable to most identification strategies commonly used in the social and health sciences. As usual, regression, weighting or doubly robust estimators can be used to implement these assumptions. There is no magic in achieving causal inference, there is simply the union of substantive assumptions and statistical approximation. Sometimes these approaches are sufficient to achieve consistent point estimates of the quantities of interest, but a healthy dose of agnosticism goes a long way.

[96]These strategies are discussed in detail and with great clarity in Angrist and Pischke (2009) and Morgan and Winship (2014).

6 Parametric Models

All models are wrong, but some are useful.

— GEORGE BOX

We have thus far focused on making *credible* inferences based on an agnostic view of the world. That is to say, our estimates, confidence intervals, and p-values have not depended on having full (or nearly full) knowledge of how our data were generated. Our assumptions have tended to be minimal and substantive. This approach has not been the historical norm in statistics or econometrics. Instead, these disciplines have primarily been developed around the idea of constructing statistical *models* to explain observed phenomena, and then imposing strong distributional assumptions to justify inferences within the framework of the model.

In this chapter, we show how such stronger distributional assumptions can be used profitably to assist in statistical inference, even when these assumptions do not exactly hold.[97]

6.1 Models and Parameters

We will consider *parametric* models, i.e. models that assume that the distribution of some outcome variable Y_i, conditional on a set of explanatory variables $\mathbf{X}_i$, is fully characterized by a finite number of *parameters*. We formalize this as follows.

Definition 6.1.1. *Parametric Model*

For a random variable Y_i and a random vector $\mathbf{X}_i$, a parametric model is an assumed conditional PMF/PDF for Y_i given $\mathbf{X}_i$,

$$f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i) = g(y_i, \mathbf{x}_i, \boldsymbol{\theta}),$$

where g is a known function and $\boldsymbol{\theta} \in \mathbb{R}^K$.

Note that $\boldsymbol{\theta}$, the *parameter vector*, is assumed to have a finite number of elements. This contrasts with our approach in earlier chapters: for all of our key results, we did not require any distributional assumptions on any of the variables—g was allowed to be completely unknown and perhaps characterized by an infinite-dimensional $\boldsymbol{\theta}$.

To illustrate how parametric models works, we'll start with an extremely simple example: the (potentially) biased coin that we considered in Example 1.1.4.

Example 6.1.1. *A Biased Coin Flip*

Let $Y_i = 0$ if the coin comes up heads and $Y_i = 1$ if the coin comes up tails. The one thing we don't know is the probability p that the coin turns up tails. Since there are no explanatory variables, we need only

[97] For those interested in a more traditional treatment of parametric models, we recommend Freedman (2009) as an exceptionally clear textbook.

consider the (unconditional) PDF $f_{Y_i}(y_i)$. Let

$$f_{Y_i}(y_i) = g(y_i, p) = \begin{cases} 1-p & : y_i = 0 \\ p & : y_i = 1 \\ 0 & : \text{otherwise} \end{cases}$$

Then $\boldsymbol{\theta} = p$, since p completely determines the distribution of the random variable Y_i. This is, technically, a parametric model for Y_i. If we knew just one number, p, then we would know everything there is to know about the distribution of Y_i. Formally, Y_i follows a *Bernoulli distribution* with parameter p.

Similarly, consider a roll of a (potentially) loaded six-sided die.

Example 6.1.2. *A Loaded Die Roll*

Let Y_i take on the value of the side of the die that comes up. Our model is:

$$f_{Y_i}(y_i) = g(y_i, \mathbf{p}) = \begin{cases} p_1 & : y_i = 1 \\ p_2 & : y_i = 2 \\ p_3 & : y_i = 3 \\ p_4 & : y_i = 4 \\ p_5 & : y_i = 5 \\ 1 - \sum_{k=1}^{5} p_k & : y_i = 6 \\ 0 & : \text{otherwise} \end{cases}$$

where $\mathbf{p} = (p_1, p_2, p_3, p_4, p_5)$. Then $\boldsymbol{\theta} = \mathbf{p}$, since those five values completely determine the distribution of Y_i. This is, again, a parametric model for Y_i. Formally, Y_i follows a *categorical distribution* (also known as a *generalized Bernoulli distribution*) with event probabilities p_1, p_2, p_3, p_4, p_5, and $1 - \sum_{k=1}^{5} p_k$.

Examples 6.1.1 and 6.1.2 represent the simplest sorts of statistical models. These models describe the (unconditional) distribution of a random variable Y_i in terms of a finite number of parameters. However, as Definition 6.1.1 suggests, some of the most important and widely-used models characterize the *conditional* distribution of an outcome variable Y_i given a set of explanatory variables $\mathbf{X}_i$. Let's now consider some of these more complicated models.

6.1.1 The Classical Linear Model

The *classical linear model* is the workhorse of conventional applied econometrics and statistics. This model can be defined in accordance with our general definition of a parametric model as follows.

Definition 6.1.2. *Classical Linear Model*

For a random variable Y_i and a random vector $\mathbf{X}_i$, the classical linear model is a parametric model with

$$g(y_i, \mathbf{x}_i, (\boldsymbol{\beta}, \sigma)) = \phi(y_i, (\mathbf{x}_i\boldsymbol{\beta}, \sigma^2))$$

where $\phi(y, (\mu, \sigma^2))$ denotes the PDF of the normal distribution with mean μ and variance σ^2.

Equivalently, the classical linear model can also be written in the following (much more common) form:

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + U_i,$$

where $U_i \sim N(0, \sigma^2)$.

Let's consider an example of the classical linear model in practice.

Example 6.1.3. *Modeling SAT Scores*

Suppose that we are trying to model SAT scores for high school students. We have the following information about each student i:

- $Income_i$: student i's family income in the previous calendar year,
- $Tutoring_i$: an indicator for whether or not student i received tutoring,
- GPA_i: student i's high school GPA, and
- Y_i: student i's SAT score.

Let $\mathbf{X}_i = (1, Income_i, Tutoring_i, GPA_i)$. Then we might impose the following linear model:

$$\begin{aligned} Y_i &= \mathbf{X}_i\boldsymbol{\beta} + U_i \\ &= \beta_0 + \beta_1 \cdot Income_i + \beta_2 \cdot Tutoring_i + \beta_3 \cdot GPA_i + U_i, \end{aligned}$$

where $U_i \sim N(0, \sigma^2)$. The final term U_i represents the "disturbance," i.e., all variation in SAT scores that is not explained by income, tutoring, GPA, and the constant.

In this example, the linear model implies that, if we knew just five numbers, we would know *everything* about the distribution of SAT scores conditional on income, tutoring, and GPA. It is a parametric model with parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$. The conditional PDF of Y_i given $\mathbf{X}_i$ is assumed to be of the form:

$$f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i) = \phi(y_i, (\beta_0 + \beta_1 \cdot Income_i + \beta_2 \cdot Tutoring_i + \beta_3 \cdot GPA_i, \sigma^2)).$$

Defenders of the classical linear model would argue that "all of the things that we don't measure probably add up to a normal distribution" and that the assumption that $U_i \sim N(0, \sigma^2)$ is a harmless technical convenience. It is true that assuming a normally distributed disturbance is not entirely unreasonable, since the CLT implies that the sum of many small, independent, random perturbations will tend to be normally distributed. But, at the same time, this assumption is almost certainly wrong.

Note that nothing about the statistical model implies any causal interpretation. Rather, it simply embeds a set of assumptions about how the explanatory variables—family income, tutoring, and GPA—characterize the conditional distribution of test scores.

6.1.2 Binary Choice Model

Let's consider another class of commonly used statistical models: *binary choice models*.

Definition 6.1.3. *Binary Choice Model*

For a binary random variable Y_i and a random vector $\mathbf{X}_i$, a binary choice model is a parametric model with

$$g(y_i, \mathbf{X}_i, \boldsymbol{\beta}) = \begin{cases} 1 - h(\mathbf{X}_i\boldsymbol{\beta}) & : y_i = 0 \\ h(\mathbf{X}_i\boldsymbol{\beta}) & : y_i = 1 \\ 0 & : \text{otherwise,} \end{cases}$$

where $h : \mathbb{R} \to [0, 1]$ is a known function.

The two most commonly used binary choice models are *logit* and *probit*, which are defined as follows.

Definition 6.1.4. *Logit Model*

For a binary random variable Y_i and a random vector $\mathbf{X}_i$, a logit model is a binary choice model where the link function is the logistic function:

$$h(\mathbf{X}_i\boldsymbol{\beta}) = \frac{e^{\mathbf{X}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{X}_i\boldsymbol{\beta}}}.$$

Definition 6.1.5. *Probit Model*

For a binary random variable Y_i and a random vector $\mathbf{X}_i$, a probit model is a binary choice model where the link function is the standard normal CDF:

$$h(\mathbf{X}_i\boldsymbol{\beta}) = \Phi(\mathbf{X}_i\boldsymbol{\beta}).$$

Figure 6.1.1 displays the graphs of these functions.

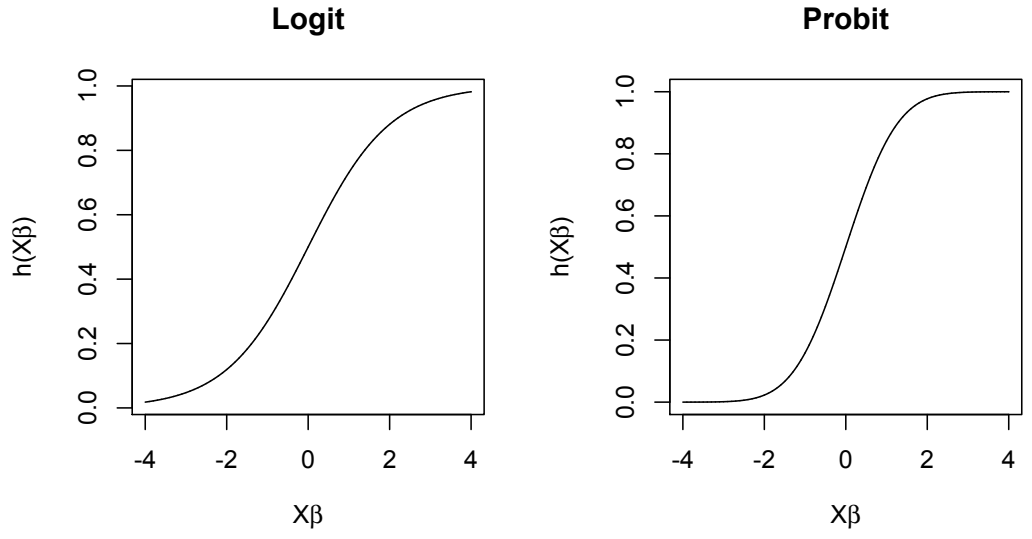


Figure 6.1.1 *Logit and Probit Functions*

Note that the link functions for logit and probit have very similar shapes, differing primarily in scale. In both models, large changes in $\mathbf{X}_i\boldsymbol{\beta}$ at extreme values correspond to relatively small changes in the conditional probability of $Y_i = 1$.

We now provide an example of how binary choice models work in practice.

Example 6.1.4. *Modeling the Decision to Buy a Car*

Suppose that we are trying to model the probability that an individual will buy a car. We have the following information about each individual i:

- $Income_i$: individual i's income in the previous calendar year,
- $Distance_i$: individual i's distance from their workplace,
- $PublicTransit_i$: an indicator for whether or not individual i has access to public transportation, and
- Y_i: an indicator for whether or not individual i bought a car.

Let $\mathbf{X}_i = (1, Income_i, Distance_i, PublicTransit_i)$. Then we might assume that

$$\begin{aligned}\Pr(Y_i = 1|\mathbf{X}_i) &= h(\mathbf{X}_i\boldsymbol{\beta}) \\ &= h(\beta_0 + \beta_1 \cdot Income_i + \beta_2 \cdot Distance_i + \beta_3 \cdot PublicTransit_i).\end{aligned}$$

Once we assume the form of h, this is a parametric model: assuming the model holds, knowledge of $\boldsymbol{\beta}$ tells us everything about the conditional distribution of Y_i given $\mathbf{X}_i$. For example, if h is the standard normal CDF, then we have a probit model with

$$\begin{aligned}\Pr(Y_i = 1|\mathbf{X}_i) &= \Phi(\mathbf{X}_i\boldsymbol{\beta}) \\ &= \Phi(\beta_0 + \beta_1 \cdot Income_i + \beta_2 \cdot Distance_i + \beta_3 \cdot PublicTransit_i).\end{aligned}$$

If β_0, β_1, β_2, and β_3 are known, this equation fully describes the conditional distribution of Y_i given $\mathbf{X}_i$.

Many people believe that, when working with binary outcome data, it is essential to use a binary choice model instead of the classical linear model. If the goal of the research endeavor is to parametrically model a binary response, then certainly the classical linear model must be wrong. Logit and probit are often more sensible options for modeling, but there is certainly no guarantee that it will approximate $f_{Y_i|\mathbf{X}_i}(y)$ better for any given generative process.

6.2 Maximum Likelihood Estimation

Once we have chosen a model to describe an observed phenomenon, how can we estimate its parameters from the data? As usual, we will assume that we have i.i.d. observations. It can then be shown that, if we

assume that the parametric model is true, then a procedure known as *maximum likelihood (ML) estimation* yields consistent, asympotically efficient, and asymptotically normal estimates of the parameters. Indeed, under this assumption, there exists no "regular" estimator that has better asymptotic efficiency.

ML estimation is a type of *M-estimation*. M-estimators are a broad class of estimators that are computed by minimizing a *loss function* given the observed data.[98] OLS is a classic example of an M-estimator, where the loss function is the sum of squared residuals.

6.2.1 The Logic of Maximum Likelihood Estimation

For ease of exposition, we'll illustrate the formalization of ML estimation for the case of discrete random variables. The logic is the same for continuous random variables but requires working directly with density functions rather than event probabilities.

Suppose that we have n i.i.d. observations of a discrete random vector $(Y_i, \mathbf{X}_i)$, and suppose that we assume the model

$$f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i) = g(y_i, \mathbf{x}_i, \boldsymbol{\theta}),$$

so that the conditional PMF of Y_i given $\mathbf{X}_i$ is entirely characterized by some known function g and the parameter vector $\boldsymbol{\theta}$. The idea behind ML estimation is that we want to find the parameters $\boldsymbol{\theta}$ that would have maximized the probability of obtaining the data that we actually observed. In other words, since we don't know the parameters but we *do* know the outcome—or rather, n i.i.d. outcomes—our best guess for the parameters is the values that would have made that particular set of outcomes as likely as possible.

To illustrate the logic behind ML estimation, suppose that we were to consider $\boldsymbol{\theta}$ to be a random variable.[99] Suppose we had a single observation of $(Y_i, \mathbf{X}_i)$. Then, intuitively, we would want to choose an estimate $\hat{\boldsymbol{\theta}}$ that would maximize the probability that $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ conditional on the observed outcome for $(Y_i, \mathbf{X}_i)$. By Bayes' Rule:

$$\Pr(\boldsymbol{\theta}|Y_i, \mathbf{X}_i) = \frac{\Pr(Y_i, \mathbf{X}_i|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta})}{\Pr(Y_i, \mathbf{X}_i)}.$$[100]

By the Multiplicative Law of Probability (conditional on $\boldsymbol{\theta}$),

$$\Pr(Y_i, \mathbf{X}_i|\boldsymbol{\theta}) = \Pr(Y_i|\mathbf{X}_i, \boldsymbol{\theta}) \Pr(\mathbf{X}_i|\boldsymbol{\theta}),$$

so we have:

$$\Pr(\boldsymbol{\theta}|Y_i, \mathbf{X}_i) = \frac{\Pr(Y_i|\mathbf{X}_i, \boldsymbol{\theta}) \Pr(\mathbf{X}_i|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta})}{\Pr(Y_i, \mathbf{X}_i)}.$$

[98] Maximum likelihood estimation, as the name suggests, actually involves *maximizing* an objective function, as we will see in Section 6.2.1. But maximizing an objective function is equivalent to minimizing the negative of that function, so ML estimation can be considered a type of M-estimation.

[99] Here we are effectively departing briefly from our otherwise strict adherence to the frequentist paradigm and entering the realm of Bayesian statistics. This is the easiest way to convey the intuition behind ML estimation.

[100] As with $\mathrm{E}[\,\cdot\,|X]$, $V(\,\cdot\,|X)$, etc., when we write a random variable by itself (e.g., X) as an argument of an operator (e.g. $\Pr(\,\cdot\,)$) where there is usually an expression like $X = x$, we are denoting the random variable that takes on the value of the operator when X takes on the value x. E.g., if $G(x) = \Pr(X = x)$, then $\Pr(X)$ denotes the random variable $Z = G(X)$; if $H(y, x) = \Pr(Y = y|X = x)$, then $\Pr(Y|X)$ denotes the random variable $W = H(Y, X)$.

Note that we have not modeled the distribution of $\mathbf{X}_i$ in terms of $\boldsymbol{\theta}$; $\boldsymbol{\theta}$ only governs the conditional distribution of Y_i given $\mathbf{X}_i$. Thus, $\mathbf{X}_i \perp\!\!\!\perp \boldsymbol{\theta}$, which implies that

$$\Pr(\mathbf{X}_i|\boldsymbol{\theta}) = \Pr(\mathbf{X}_i = x_i),$$

and so now we have:

$$\Pr(\boldsymbol{\theta}|Y_i, \mathbf{X}_i) = \frac{\Pr(Y_i|\mathbf{X}_i, \boldsymbol{\theta}) \Pr(\mathbf{X}_i) \Pr(\boldsymbol{\theta})}{\Pr(Y_i, \mathbf{X}_i)}.$$

The *likelihood function* is defined as:

$$\mathcal{L}(\boldsymbol{\theta}|Y_i, \mathbf{X}_i) = \Pr(Y_i|\mathbf{X}_i, \boldsymbol{\theta}),$$

so that

$$\begin{aligned}\Pr(\boldsymbol{\theta}|Y_i, \mathbf{X}_i) &= \frac{\mathcal{L}(\boldsymbol{\theta}|Y_i, \mathbf{X}_i) \Pr(\mathbf{X}_i) \Pr(\boldsymbol{\theta})}{\Pr(Y_i, \mathbf{X}_i)} \\ &= \mathcal{L}(\boldsymbol{\theta}|Y_i, \mathbf{X}_i) \Pr(\boldsymbol{\theta}) \cdot \frac{\Pr(\mathbf{X}_i)}{\Pr(Y_i, \mathbf{X}_i)}.\end{aligned}$$

Since the last term is constant with respect to $\boldsymbol{\theta}$, we have:

$$\Pr(\boldsymbol{\theta}|Y_i, \mathbf{X}_i) \propto \mathcal{L}(\boldsymbol{\theta}|Y_i, \mathbf{X}_i) \Pr(\boldsymbol{\theta}).$$

So, if we want to find the $\hat{\boldsymbol{\theta}}$ that is most probable, and our *priors*, or ex ante beliefs, for $\Pr(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}})$ are completely flat, i.e., for any $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}'$, $\Pr(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}) = \Pr(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}')$, then the $\hat{\boldsymbol{\theta}}$ with the maximum likelihood is also the $\hat{\boldsymbol{\theta}}$ that is most probable. There is no better guess unless we have different priors. And, in large samples, the priors become completely irrelevant anyway.

We can now formally define the maximum likelihood estimator.

Definition 6.2.1. *Maximum Likelihood Estimator*

Let Y_i be a random variable and $\mathbf{X}_i$ be a random vector. Let $(Y_1, \mathbf{X}_1)$, $(Y_2, \mathbf{X}_2)$, ..., $(Y_n, \mathbf{X}_n)$ be i.i.d. observations of $(Y_i, \mathbf{X}_i)$, and let $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ and $\mathbb{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$. Then for the parametric model $f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i) = g(y_i, \mathbf{x}_i, \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \mathbb{R}^K$, the maximum likelihood estimator of $\boldsymbol{\theta}$ is:

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\hat{\boldsymbol{\theta}} \in \mathbb{R}^K}{\arg\max}\, \mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}|\mathbf{Y}, \mathbb{X}),$$

where $\mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}|\mathbf{Y}, \mathbb{X})) = \prod_{i=1}^{n} g(Y_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}})$.

To see how ML estimation works, let's return to our coin flip example.

Example 6.2.1. *ML Estimation for a Biased Coin Flip*

As usual, let $Y_i = 0$ if the coin comes up heads and $Y_i = 1$ if the coin comes up tails. Our model is:

$$f_{Y_i}(y_i) = g(y_i, p) = \begin{cases} 1-p & : y_i = 0 \\ p & : y_i = 1 \\ 0 & : \text{otherwise} \end{cases}$$

The one thing we don't know is the parameter p, the probability that the coin turns up tails. Suppose that we observed n i.i.d. coin flips: $\mathbf{y} = (y_1, y_2, ..., y_n)$. How do we compute the maximum likelihood estimate of p? First note that, for $y_i \in \{0, 1\}$,

$$\begin{aligned} g(y_i, \hat{p}) &= \begin{cases} 1-\hat{p} & : y_i = 0 \\ \hat{p} & : y_i = 1 \end{cases} \\ &= (1-\hat{p})^{1-y_i} \hat{p}^{y_i}. \end{aligned}$$

And since we have i.i.d. observations, it follows that

$$\Pr(\mathbf{Y} = \mathbf{y} | p = \hat{p}) = \prod_{i=1}^{n} \Pr(Y_i = y_i | p = \hat{p}).$$

Thus,

$$\begin{aligned} \mathcal{L}(p = \hat{p} | \mathbf{Y} = \mathbf{y}) &= \Pr(\mathbf{Y} = \mathbf{y} | p = \hat{p}) \\ &= \prod_{i=1}^{n} \Pr(Y_i = y_i | p = \hat{p}) \\ &= \prod_{i=1}^{n} g(y_i, \hat{p}) \\ &= \prod_{i=1}^{n} (1-\hat{p})^{1-y_i} \hat{p}^{y_i} \\ &= (1-\hat{p})^{\left[\sum_{i=1}^{n}(1-y_i)\right]} \hat{p}^{\left[\sum_{i=1}^{n} y_i\right]} \\ &= (1-\hat{p})^{N_0} \hat{p}^{N_1}, \end{aligned}$$

where N_0 is the number of heads, and N_1 is the number of tails.[101]

Thus, to compute the ML estimate for p, we want to find the $\hat{p}$ that maximizes

$$\mathcal{L}(p = \hat{p} | \mathbf{Y} = \mathbf{y}) = (1-\hat{p})^{N_0} \hat{p}^{N_1}$$

given the observed data $\mathbf{y} = (y_1, y_2, ..., y_n)$. Applying the first order condition,

$$\frac{\partial}{\partial \hat{p}} \mathcal{L}(p = \hat{p} | \mathbf{Y} = \mathbf{y}) = 0.$$

[101] We use uppercase letters here since N_0 and N_1 are random variables.

So, applying the product rule, we have

$$
\begin{aligned}
-N_0(1-\hat{p})^{N_0-1}\hat{p}^{N_1} + N_1(1-\hat{p})^{N_0}\hat{p}^{N_1-1} &= 0 \\
N_1(1-\hat{p})^{N_0}\hat{p}^{N_1-1} &= N_0(1-\hat{p})^{N_0-1}\hat{p}^{N_1} \\
\frac{(1-\hat{p})^{N_0}\hat{p}^{N_1-1}}{(1-\hat{p})^{N_0-1}\hat{p}^{N_1}} &= \frac{N_0}{N_1} \\
\frac{1-\hat{p}}{\hat{p}} &= \frac{N_0}{N_1} \\
\frac{1}{\hat{p}} - 1 &= \frac{N_0}{N_1} \\
\frac{1}{\hat{p}} &= \frac{N_0+N_1}{N_1} \\
\frac{1}{\hat{p}} &= \frac{n}{N_1} \\
\hat{p} &= \frac{N_1}{n}.
\end{aligned}
$$

Thus, the maximum likelihood estimate of the probability of the coin coming up tails is the proportion of flips that came up tails. That's just the usual plug-in estimate! We know from the WLLN that N_1/n is a consistent estimator for p.

To illustrate this more concretely, suppose that we tossed the coin $n = 3$ times and observed $\mathbf{y} = (1, 0, 1)$. What would be the maximum likelihood estimate for p? Figure 6.2.1 shows the likelihood as a function of p given these data.

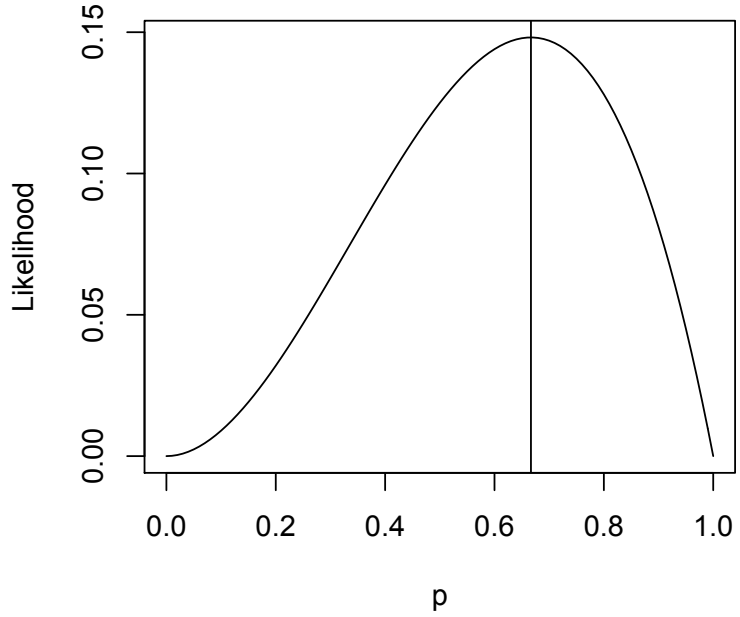


Figure 6.2.1 *A Likelihood Function for Three Coin Tosses*

Thus, the maximum likelihood estimate for p is $\hat{p}_{ML} = 2/3$, which is again simply the plug-in estimate, i.e. the proportion of flips that came up tails.

Usually, it's more convenient to work with the *log-likelihood* rather than the likelihood. Because the natural logarithm—which we write simply as $\log$—is a strictly increasing function, the maximum log-likelihood must be attained at the same point as the maximum likelihood. Thus,

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\hat{\boldsymbol{\theta}} \in \mathbb{R}^K}{\arg\max} \left[\log \mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} | \mathbf{Y}, \mathbb{X}) \right].$$

Transforming the problem in this way confers a number of benefits, including 1) transforming products into sums, which makes taking the derivative easier, and 2) preventing numerical instability with large n. Let's see how this works in the case of the classical linear model.

Example 6.2.2. *ML Estimation of the Classical Linear Model*

Let Y_i be a random variable and $\mathbf{X}_i$ be a random vector. Suppose that

$$Y_i = \mathbf{X}_i \boldsymbol{\beta} + U_i,$$

where $U_i \sim N(0, \sigma^2)$ conditional on $\mathbf{X}_i$. Let $(Y_1, \mathbf{X}_1)$, $(Y_2, \mathbf{X}_2)$, ..., $(Y_n, \mathbf{X}_n)$ be i.i.d. observations of

$(Y_i, \mathbf{X}_i)$, and let $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ and $\mathbb{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$. Then

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}|\mathbf{Y}, \mathbb{X}) &= \prod_{i=1}^{n} \phi\left(Y_i, (\mathbf{X}_i\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)\right) \\ &= \prod_{i=1}^{n} \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{\frac{-(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2}{2\hat{\sigma}^2}}.\end{aligned}$$

So the log-likelihood is:

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}|\mathbf{Y}, \mathbb{X}) &= \log\left[\prod_{i=1}^{n} \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{\frac{-(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2}{2\hat{\sigma}^2}}\right] \\ &= \sum_{i=1}^{n} \log\left[\frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{\frac{-(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2}{2\hat{\sigma}^2}}\right] \\ &= \sum_{i=1}^{n} \left[\log\left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right) + \frac{-(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2}{2\hat{\sigma}^2}\right].\end{aligned}$$

We want to find the $\hat{\boldsymbol{\beta}}$ that maximizes $\log \mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}|\mathbf{Y}, \mathbb{X})$. (We don't really care about estimating σ.)

$$\begin{aligned}\underset{\hat{\boldsymbol{\beta}} \in \mathbb{R}^K}{\arg\max} \left[\log \mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}|\mathbf{Y}, \mathbb{X})\right] &= \underset{\hat{\boldsymbol{\beta}} \in \mathbb{R}^K}{\arg\max} \sum_{i=1}^{n} \left[\log\left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right) + \frac{-(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2}{2\hat{\sigma}^2}\right] \\ &= \underset{\hat{\boldsymbol{\beta}} \in \mathbb{R}^K}{\arg\max} \sum_{i=1}^{n} \left[-(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2\right] \\ &= \underset{\hat{\boldsymbol{\beta}} \in \mathbb{R}^K}{\arg\min} \sum_{i=1}^{n} \left[(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2\right].\end{aligned}$$

So the maximum likelihood estimator for $\boldsymbol{\beta}$ is the $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared residuals. That's just the OLS estimator!

Thus, the ML estimator for $\boldsymbol{\beta}$ is the regression estimator when it is assumed that:

- the linear model holds, and
- the errors are $U_i \sim N(0, \sigma^2)$ for every observation.

Many textbooks justify OLS under these strong distributional assumptions, but these assumptions are not necessary. Contrast this with our nonparametric results, which require no such distributional assumptions. Under i.i.d. sampling, OLS consistently estimates the BLP of the outcome (and the BLP of the CEF).

It is an unusual feature of the classical linear model that there exists a closed-form solution for the ML estimator of $\boldsymbol{\beta}$. In practice, ML estimators rarely have closed-form solutions. Instead, one usually has to calculate the solution numerically using a computer.[102] This is the case, for example, for logit and probit.

Example 6.2.3. *Maximum Likelihood Estimation of the Probit Model*

Let Y_i be a binary random variable and $\mathbf{X}_i$ be a random vector. Suppose that

$$\Pr(Y_i = 1|\mathbf{X}_i) = \Phi(\mathbf{X}_i\boldsymbol{\beta}),$$

which implies that

$$f_{Y_i|\mathbf{X}_i}(Y_i|\mathbf{X}_i) = g(Y_i, X_i, \boldsymbol{\beta}) = \Phi(\mathbf{X}_i\boldsymbol{\beta})^{Y_i}[1 - \Phi(\mathbf{X}_i\boldsymbol{\beta})]^{1-Y_i}.$$

Let $(Y_1, \mathbf{X}_1)$, $(Y_2, \mathbf{X}_2)$, ..., $(Y_n, \mathbf{X}_n)$ be i.i.d. observations of $(Y_i, \mathbf{X}_i)$, and let $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ and $\mathbb{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$. Then

$$\mathcal{L}(\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}|\mathbf{Y}, \mathbb{X}) = \prod_{i=1}^{n} \Phi(\mathbf{X}_i\boldsymbol{\beta})^{Y_i}[1 - \Phi(\mathbf{X}_i\boldsymbol{\beta})]^{1-Y_i}.$$

Thus, the log-likelihood is:

$$\begin{aligned}
\log \mathcal{L}(\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}|\mathbf{Y}, \mathbb{X}) &= \log \left(\prod_{i=1}^{n} \Phi(\mathbf{X}_i\boldsymbol{\beta})^{Y_i}[1 - \Phi(\mathbf{X}_i\boldsymbol{\beta})]^{1-Y_i} \right) \\
&= \sum_{i=1}^{n} \log \left(\Phi(\mathbf{X}_i\boldsymbol{\beta})^{Y_i}[1 - \Phi(\mathbf{X}_i\boldsymbol{\beta})]^{1-Y_i} \right) \\
&= \sum_{i=1}^{n} \Big(Y_i \log \Phi(\mathbf{X}_i\boldsymbol{\beta}) + (1 - Y_i) \log[1 - \Phi(\mathbf{X}_i\boldsymbol{\beta})] \Big).
\end{aligned}$$

So, given the outcomes for $\mathbf{Y}$ and $\mathbb{X}$, we would obtain the ML estimate by using a computer to calculate

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\arg\max} \sum_{i=1}^{n} \Big(Y_i \log \Phi(\mathbf{X}_i\boldsymbol{\beta}) + (1 - Y_i) \log[1 - \Phi(\mathbf{X}_i\boldsymbol{\beta})] \Big).$$

6.2.2 Properties of Maximum Likelihood Estimation

We now provide a loose proof of the consistency of ML estimation assuming that the parametric model is correct. It will draw on the idea of the plug-in principle.

[102] There are many algorithms that can be implemented to find the minimum of a function, generally involving some sort of iterated search. For example, the *Newton-Raphson method* works by taking the second-order polynomial approximation (i.e., Taylor expansion) of the function at a given point by taking numerical derivatives. The first point must be specified as a starting (or "seed") value. Then it updates to guess the point that maximizes the polynomial approximation. Then it repeats the procedure until it finds a fixed point—a point where the estimate of the maximum doesn't change by more than a specified tolerance when the procedure is repeated again. Note: depending on the seed value, this method, and many others like it, may not necessarily converge to the global maximum if the function is not strictly concave.

Theorem 6.2.1. *Consistency of the Maximum Likelihood Estimator*

Let Y_i be a random variable and $\mathbf{X}_i$ be a random vector. Assume that the parametric model $f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i) = g(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0)$ holds (where $\boldsymbol{\theta}_0$ denotes the true value of $\boldsymbol{\theta}$) and that the likelihood function satisfies some requisite regularity conditions (e.g., smoothness, compactness, unique maximum). Then $\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{p} \boldsymbol{\theta}$.

Proof: We provide an illustrative proof sketch. For a more complete derivation, see Wasserman (2004).

Let $(Y_1, \mathbf{X}_1)$, $(Y_2, \mathbf{X}_2)$, ..., $(Y_n, \mathbf{X}_n)$ be i.i.d. observations of $(Y_i, \mathbf{X}_i)$, and let $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ and $\mathbb{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$. Let

$$L_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n} \log \mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} | \mathbf{Y}, \mathbb{X}).$$

By definition (and the assumption of a unique maximum), $\hat{\boldsymbol{\theta}}_{ML}$ uniquely maximizes $\log \mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} | \mathbf{Y}, \mathbb{X})$ and therefore uniquely maximizes $L_n(\hat{\boldsymbol{\theta}})$. Now, note that

$$\frac{1}{n} \log \mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} | \mathbf{Y}, \mathbb{X}) = \frac{1}{n} \log \left[\prod_{i=1}^{n} g(Y_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}) \right] = \frac{1}{n} \sum_{i=1}^{n} \log g(Y_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}).$$

This is a sample mean, so by the WLLN,

$$\frac{1}{n} \sum_{i=1}^{n} \log g(Y_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathrm{E}\left[\log g(Y_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}) \right]$$

Let $L(\hat{\boldsymbol{\theta}}) = \mathrm{E}\left[\log g(Y_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}) \right] = \mathrm{E}\left[\log \mathcal{L}(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} | Y_i, \mathbf{X}_i) \right]$, so that we can write the above statement as

$$L_n(\hat{\boldsymbol{\theta}}) \xrightarrow{p} L(\hat{\boldsymbol{\theta}}),$$

i.e., the sample mean of the likelihood of each observation given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ converges to the expected likelihood of a single observation given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

We now require the following lemma: $\forall \hat{\boldsymbol{\theta}}$,

$$L(\hat{\boldsymbol{\theta}}) \leq L(\boldsymbol{\theta}_0),$$

with

$$L(\hat{\boldsymbol{\theta}}) = L(\boldsymbol{\theta}_0) \iff \Pr\left[g(Y_i, \mathbf{X}, \hat{\boldsymbol{\theta}}) = g(Y_i, \mathbf{X}_i, \boldsymbol{\theta}_0) \right] = 1.$$

I.e., the expected value of the log-likelihood function is uniquely maximized when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. We omit the proof of this lemma, but intuitively: for any $\hat{\boldsymbol{\theta}}$ that *wasn't* the true $\boldsymbol{\theta}_0$, we know that $\boldsymbol{\theta}_0$ would have to fit the true distribution better.

So, to complete our sketch, we have shown that

- $\hat{\boldsymbol{\theta}}_{ML}$ uniquely maximizes $L_n(\hat{\boldsymbol{\theta}})$,
- $L_n(\hat{\boldsymbol{\theta}}) \xrightarrow{p} L(\hat{\boldsymbol{\theta}})$, and

- $\boldsymbol{\theta}_0$ uniquely maximizes $L(\hat{\boldsymbol{\theta}})$.

Thus, the ML estimate $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, so $\hat{\boldsymbol{\theta}}$ is consistent. $\square$

What this proof sketch illustrates is that the ML estimator is itself a type of plug-in estimator. The inferential target, $\boldsymbol{\theta}_0$, is a feature of the population distribution: the parameter values that maximize the expected value of the log-likelihood function. To estimate this population quantity, we simply use the sample analog: the parameter values that maximize the sample mean of the likelihood function. Since the sample will more and more closely approximate the population distribution as n grows large, the ML estimate will get closer and closer to the true parameter values.

Furthermore, under proper specification, ML estimation is also asymptotically normal and efficient—if the parametric model is right, then with large n, there exists no better (i.e., lower MSE) estimator of $\boldsymbol{\theta}_0$. This makes sense: ML exploits *all* of the information available about the distribution. So it's not as though we're leaving anything on the table; asymptotically, we can't do any better by introducing more information from the observed data, since no more information exists.

6.2.3 Maximum Likelihood Estimation under Misspecification

What happens if the parametric model is wrong (as it almost surely is)? What, then, does ML estimation give us, asymptotically? The somewhat technical answer is that ML estimation will consistently estimate the parameters that minimize the *Kullback-Leibler (KL) divergence* of the hypothesized distribution from the true distribution.[103]

Definition 6.2.2. *Kullback-Leibler Divergence for Continuous Distributions*

For probability density functions $p(\cdot)$ and $q(\cdot)$, the Kullback-Leibler divergence,

$$KL[p(\cdot)||q(\cdot)] = \int_{-\infty}^{\infty} p(t) \log \frac{p(t)}{q(t)} dt.$$

The KL distance is defined analogously for discrete random variables, with a sum replacing the integral and PMFs replacing PDFs. Note that the KL distance between any distributions $p(\cdot)$ and $q(\cdot)$ is 0 whenever P and Q have identical distributions, as $p(t)/q(t) = 1$ at every point t. For any distributions $p(\cdot)$ and $q(\cdot)$ that differ, the KL distance will be nonzero and positive.

Theorem 6.2.2. *The ML Estimator is Consistent for the MKLD Predictor*

The maximum likelihood estimate of $f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i)$, $g(y_i, \mathbf{x}_i, \hat{\boldsymbol{\theta}}_{ML})$, consistently estimates the minimum Kullback-Leibler divergence (MKLD) approximation of $f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i)$ among all probability distributions of the form $g(y_i, \mathbf{x}_i, \boldsymbol{\theta})$. I.e., assuming that $\tilde{\boldsymbol{\theta}}_0 = \arg\min_{\hat{\boldsymbol{\theta}} \in \mathbb{R}^K} KL[f_{Y_i|\mathbf{X}_i}(\cdot|\mathbf{x}_i)||g(\cdot, \mathbf{x}_i, \hat{\boldsymbol{\theta}})]$ is unique, then

$$\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{p} \tilde{\boldsymbol{\theta}}_0.$$

[103] The parameters that minimize KL divergence are sometimes referred to as the *quasi-ML* parameters.

We omit the proof of this theorem. What's important here is to note that this fact resembles a result we derived for OLS, only with a different metric for the "distance" between distributions (KL divergence rather than mean squared error). Among all linear predictors, OLS consistently estimates the one "closest" (in terms of mean squared error) to the CEF. Likewise, among all conditional distributions of the form given by the hypothesized model, ML estimation consistently estimates the one "closest" (in terms of KL divergence) to the true conditional distribution. In short, there exists a sensible definition of "best" under which ML estimation consistently estimates the best-fitting distribution among all distributions in the assumed parametric family.

The idea behind the proof of Theorem 6.2.2 is again related to the plug-in principle. There is a population quantity of interest, $\tilde{\boldsymbol{\theta}}_0$: the parameter values that minimize the KL divergence of the hypothesized distribution from the true population distribution. It can be shown that the ML estimate is equal to the sample analog: the parameter values that minimize the KL divergence of the hypothesized distribution from the observed data. So since the distribution of the observed data converges to the population distribution as n grows large, the ML estimate converges to the value that minimizes the KL divergence of the hypothesized distribution from the true population distribution.

Note that if in fact $f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i) = g(y_i, \mathbf{x}_i, \boldsymbol{\theta}_0)$, then $\text{KL}[f_{Y_i|\mathbf{X}_i}(\cdot|\mathbf{x}_i)||g(\cdot, \mathbf{x}_i, \boldsymbol{\theta}_0)] = 0$. Thus if the model is correct, the MKLD approximation is the true distribution, so $\tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$, and thus ML estimation consistently estimates the true parameter values, as before. Under proper specification, ML asymptotically gets the distribution exactly right. Under misspecification, ML asymptotically gets it "as right as possible." Let's consider a visualization.

Example 6.2.4. *Maximum Likelihood of an Unusual Distribution under Misspecification*

Suppose that the true distribution of Y_i were given by the histogram on the lefthand side of Figure 6.2.2. Suppose that we erroneously assumed that Y_i was normally distributed:

$$g(y_i, (\mu, \sigma)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}.$$

Then the asymptotic ML estimate of the distribution would be as shown in red on the righthand side of the figure.

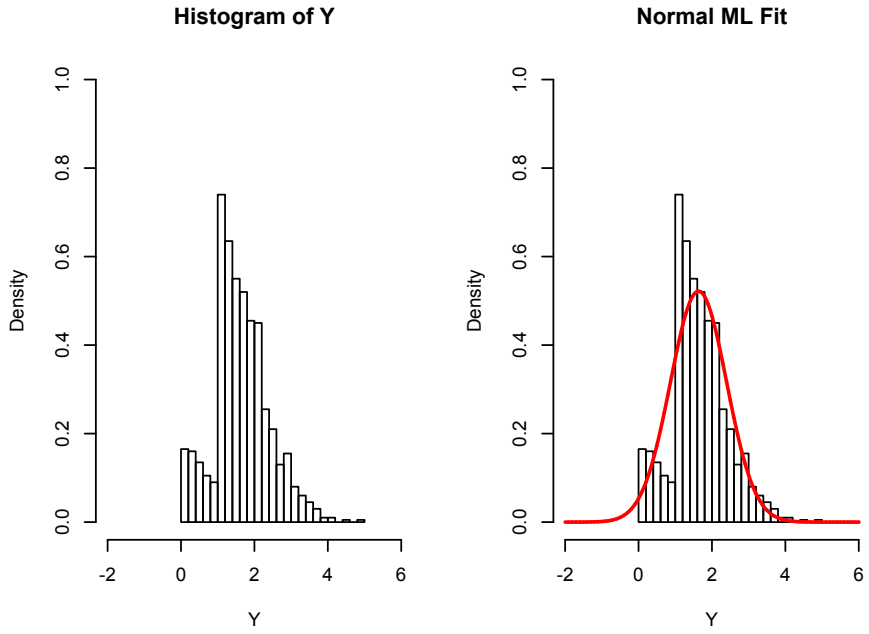


Figure 6.2.2 *MKLD Approximation of an Unusual Distribution*

A simpler example is the coin flip from Example 6.1.1. When we observe $\mathbf{y} = (1, 0, 1)$, we can match that empirical distribution exactly by setting $p = 2/3$. With i.i.d. observations, the empirical distribution will converge to the true distribution as n grows large. Thus, asymptotically, the ML estimate will exactly characterize the probability distribution of Y_i.

Cases like the coinflip, where $f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i)$ can be ensured to be properly specified, may be rare in practice, however. In general, ML estimation is best thought of as an explicit means for approximating $f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i)$.

6.2.4 Plug-in Estimation and the Conditional Expectation Function

Once we've started thinking of maximum likelihood as an approximation tool, we can start to see its real benefits. Suppose that what we're interested in is not the parameters of the model per se, but instead some other feature of the distribution of $(Y_i, \mathbf{X}_i)$. We know from Chapter 2 that the joint CDF of Y_i and $\mathbf{X}_i$ is nonparametrically identified. But we can use ML estimation to potentially improve the precision of our estimates by imposing a parametric model that seems likely to approximate the shape of the true CEF reasonably well.

We will begin by assuming that the parametric model is correct; i.e., $f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i) = g(Y_i, \mathbf{X}_i, \boldsymbol{\theta})$. When this assumption only holds approximately, then our results too will represent approximations. Let us consider the CEF of Y_i given $\mathbf{X}_i$. Given the definition of a conditional expectation (Definition 1.4.11),

$$\mathrm{E}\left[Y_i|\mathbf{X}_i\right] = \int_{y_i \in \mathrm{Supp}\left(Y_i\right)} y g(y_i, \mathbf{X}_i, \boldsymbol{\theta}) dy$$

A natural plug-in estimator would be substituting the ML estimate, $\hat{\boldsymbol{\theta}}_{ML}$, for the true values of $\boldsymbol{\theta}$:

$$\hat{\mathrm{E}}\left[Y_i|\mathbf{X}_i\right] = \int_{y_i \in \mathrm{Supp}\left(Y_i\right)} y_i g(y_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}_{ML}) dy$$

If the parametric model is correct, then naturally $\hat{\mathrm{E}}\left[Y_i|\mathbf{X}_i = \mathbf{x}_i\right]$ will be consistent for the CEF evaluated at $\mathbf{X}_i = \mathbf{x}_i$.

We may also be interested in characterizing the partial derivatives of the CEF using a parametric model. Partial derivatives of the CEF are referred to as *marginal effects* in the econometrics literature, but we refrain from using the word "effect" outside the domain of a causal model. If you are so inclined, you may replace every future instance of the term "partial derivative" in this section with "marginal effect" in your mind, but bear in mind that these do not necessarily represent causal effects without further assumptions (see Section 6.2.5).

Let us assume that the CEF is continuous and once differentiable with respect to a given variable X_{ki}. Then, the partial derivative of the CEF with respect to X_{ki},

$$\left.\frac{\partial \mathrm{E}\left[Y_i|\mathbf{X}_i\right]}{\partial X_{ki}}\right|_{\mathbf{X}_i = \mathbf{x}_i}.$$

Again, a plug-in estimator using the estimated CEF,

$$\left.\frac{\partial \hat{\mathrm{E}}\left[Y_i|\mathbf{X}_i\right]}{\partial X_{ki}}\right|_{\mathbf{X}_i = \mathbf{x}_i},$$

will be consistent under proper specification.

Thus, with a parametric model, it is straightforward to characterize and estimate the CEF of Y_i given $\mathbf{X}_i$. Often, though, researchers are interested in summarizing the relationship between Y_i and a given variable X_{ki}. The two most common quantites are the *average partial derivative* and the *partial derivative at the average*. We consider these in turn.

The average partial derivative represents, how much a change in a particular variable X_{ki} moves the CEF, averaging over the entire distribution of the data. In this sense, it is the average slope of the CEF with respect to X_{ki}.

Definition 6.2.3. *Average Partial Derivative*

The average partial derivative with respect to X_{ki},

$$APD_{X_{ki}} = \mathrm{E}_{\mathbf{X}_i}\left[\left.\frac{\partial \mathrm{E}\left[Y_i|\mathbf{X}_i\right]}{\partial X_{ki}}\right|_{\mathbf{X}_i}\right],$$

where $\mathrm{E}_{\mathbf{X}_i}[\,\cdot\,]$ integrates over the distribution of $\mathbf{X}_i$.

A natural plug-in estimator is available for $APD_{X_{ki}}$: replacing the expected value with a sample mean and the estimated CEF for the CEF. The plug-in estimate of the average partial derivative with respect to X_{ki},

$$\widehat{APD_{X_{ki}}} = \frac{1}{n}\sum_{i=1}^{n}\left[\left.\frac{\partial \hat{\mathrm{E}}\left[Y_i|\mathbf{X}_i\right]}{\partial X_{ki}}\right|_{\mathbf{X}_i=\mathbf{x}_i}\right].$$

Put simply, $\widehat{APD_{X_{ki}}}$ entails estimating the partial derivative for every observation, and taking an average across all observations. If the parametric model is correct, then $\widehat{APD_{X_{ki}}}$ is a consistent estimator of $APD_{X_{ki}}$.

The partial derivative at the average represents a conceptually different target: how much a change in a particular variable X_{ki} moves the CEF, when all variables in $\mathbf{X}_i$ are equal to their expected value. It is the slope of the CEF with respect to X_{ki} at a single point.

Definition 6.2.4. *Partial Derivative at the Average*

The partial derivative with respect to X_{ki} at the average,

$$PDA_{X_{ki}} = \left.\frac{\partial \mathrm{E}\left[Y_i|\mathbf{X}_i\right]}{\partial X_{ki}}\right|_{\mathbf{X}_i=\mathrm{E}\left[\mathbf{X}_i\right]}.$$

Again, a natural plug-in estimator is available for $APD_{X_{ki}}$, replacing the expected values with sample means, and the estimated CEF for the CEF The plug-in estimate of the partial derivative at the average with respect to X_{ki},

$$\widehat{PDA_{X_{ki}}} = \left.\frac{\partial \hat{\mathrm{E}}\left[Y_i|\mathbf{X}_i\right]}{\partial X_{ki}}\right|_{\mathbf{X}_i=\overline{\mathbf{X}_i}}.$$

Note that $\widehat{PDA_{X_{ki}}}$ is easier to compute than $\widehat{APD_{X_{ki}}}$. $\widehat{PDA_{X_{ki}}}$ only requires computing the partial derivative at a single point, whereas $\widehat{APD_{X_{ki}}}$ requires computing the partial derivative at all n points in the data. However, $APD_{X_{ki}}$ is not generally equal to $PDA_{X_{ki}}$.

Example 6.2.5. *Differences between the Average Partial Derivative and the Partial Derivative at the Average*

Suppose $X_{1i} \perp\!\!\!\perp X_{2i}$ with $X_{1i} \sim U(0,1)$ and $X_{2i} \sim U(0,1)$ Further suppose that the CEF,

$$\mathrm{E}\left[Y_i|X_{1i},X_{2i}\right] = \begin{cases} -X_{1i} & : X_{2i} \leq .8 \\ 9X_{1i} & : X_{2i} > .8 \end{cases}.$$

Then by the Law of Iterated Expectations,

$$APD_{X_{1i}} = \frac{\partial(-X_{1i})}{\partial X_{1i}}\Pr[X_{2i} \leq .8] + \frac{\partial(9X_{1i})}{\partial X_{1i}}\Pr[X_{2i} > .8] = -1 \times .8 + 10 \times .2 = 1.$$

But the partial derivative at the average gives a very different result:

$$PDA_{X_{ki}} = \frac{\partial(-X_{1i})}{\partial X_{1i}} = -1.$$

While, on average, the partial derivative with respect to X_{1i} is positive, when evaluated at the average ($X_{1i} = X_{2i} = 0.5$), the partial derivative is negative.

The average partial derivative makes a lot of sense as a way to summarize the role of a particular variable in predicting Y_i—it represents how much, on average, a variable influences the CEF. While the partial derivative at the average also has a well-defined meaning, it is limited to just one particular point on the probability distribution of the explanatory variables. Thus, for a researcher interested in characterizing the overall relationship between Y_i and X_{ki}, the average partial derivative is typically preferable.

6.2.5 Causal Inference with Parametric Models

It is possible to wed the causal identification assumptions from Chapter 5 with the parametric assumptions that we've used here. Suppose that we have conditional independence of potential outcomes $Y_i(d)$ and D_i, so that, $\forall d \in \text{Supp}\,(D_i)$, $Y_i(d) \perp\!\!\!\perp D_i|\mathbf{X}_i$, and the usual positivity assumption. Further suppose that the parametric model is correct.

Then it is the case that all average potential outcomes over the support of D_i are point identified, and can be written in terms of the parametric model. The average potential outcome when $D_i = d$,

$$\begin{aligned}
\text{E}\left[Y_i(d)\right] &= \int_{\mathbf{x}_i \in \text{Supp}\,(\mathbf{X}_i)} \text{E}\left[Y_i|D_i = d, \mathbf{X}_i = \mathbf{x}_i\right] f_{\mathbf{X}_i}(\mathbf{x}_i) d\mathbf{x}_i \\
&= \int_{\mathbf{x}_i \in \text{Supp}\,(\mathbf{X}_i)} \left[\int_{y_i \in \text{Supp}\,(Y_i)} y_i g(y_i, (d, \mathbf{x}_i), \boldsymbol{\theta}) dy_i \right] f_{\mathbf{X}_i}(\mathbf{x}_i) d\mathbf{x}_i,
\end{aligned}$$

where the first equality holds due to ignorability, and the second holds due to proper specification of the parametric model.

A natural plug-in estimator for $\text{E}\left[Y_i(d)\right]$ is available by substituting $\hat{\boldsymbol{\theta}}_{ML}$ for $\boldsymbol{\theta}$ and replacing the expected value with a sample mean:

$$\hat{\text{E}}\left[Y_i(d)\right] = \frac{1}{n} \sum_{i=1}^{n} \int_{y \in \text{Supp}\,(Y_i)} y g(y, D_i = d, \mathbf{X}_i), \hat{\boldsymbol{\theta}}) dy.$$

$\hat{\text{E}}\left[Y_i(d)\right]$ is consistent for $\text{E}\left[Y_i(d)\right]$ under conditional independence, positivity and proper specification. Note that for any $d, d' \in \text{Supp}\,(D_i)$, the average causal effect of moving from d to d' may also be estimated consistently, with $\hat{\text{E}}\left[Y_i(d)\right] - \hat{\text{E}}\left[Y_i(d')\right]$.

These definitions and estimators are useful in characterizing the effect of moving from one level of the treatment to another. But we might also be interested in summarizing the entire distribution of causal effects. In particular, we may seek to characterize how much, on average, the changes in the treatment cause changes in the outcome. Assume again that the CEF is continuous and once differentiable with respect to D_i, but also that D_i has compact support.

Definition 6.2.5. *Average Marginal Causal Effect*

The average marginal causal effect of D_i is

$$AMCE = \mathrm{E}_{D_i,\mathbf{X}_i}\left[\left.\frac{\partial \mathrm{E}\left[Y_i(d)|D_i,\mathbf{X}_i\right]}{\partial D_i}\right|_{D_i,\mathbf{X}_i}\right].$$

The Average Marginal Causal Effect is very simple to interpret: averaging over the entire distribution of D_i and $\mathbf{X}_i$, how much would we expect a small change in D_i to affect Y_i? Furthermore, under the conditional independence assumption, the CEF is causal so:

$$AMCE = APD_{D_i} = \mathrm{E}_{D_i,\mathbf{X}_i}\left[\left.\frac{\partial \mathrm{E}\left[Y_i|D_i,\mathbf{X}_i\right]}{\partial D_i}\right|_{D_i,\mathbf{X}_i}\right].$$

Or, put another way, if we have conditional independence of treatment and potential outcomes, the Average Marginal Causal Effect is identical to the average partial derivative with respect to D_i. Under these circumstances and with proper specification, $\widehat{APD_{D_i}}$ is again consistent.

When the assumed parametric model is not correct, then $\widehat{APD_{D_i}}$ will not generally be consistent for $AMCE$. But it nevertheless provides a principled approximation, one that may perform better than simply imposing a linear approximation. The key identification assumption is the conditional independence assumption; the parametric model is one of convenience. Insomuch as our parametric model yields a good approximation to the CEF, so too will $\widehat{APD_{D_i}}$ provide a good approximation to $AMCE$.

6.2.6 Mixture Models

We can use an extension of ML estimation, where we will allow for the possibility that *multiple* distributions give rise to the data. This approach is known as *mixture modeling*. While one model may fail in approximating Y_i well, many may succeed.

Suppose that the distribution of Y_i conditional on $\mathbf{X}_i$ can be represented as a weighted average of parametric models,

$$f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i) = g(y_i,\mathbf{x}_i,\boldsymbol{\theta}) = p_1 g_1(\boldsymbol{\theta}_1,\mathbf{X}_i) + p_2 g_2(\boldsymbol{\theta}_2,\mathbf{X}_i) + ... + (1 - \sum_{k=1}^{K-1} p_k) g_K(\boldsymbol{\theta}_K,\mathbf{X}_i),$$

where $\boldsymbol{\theta} = (p_1,\boldsymbol{\theta}_1,p_2,\boldsymbol{\theta}_2,...,p_{K-1},\boldsymbol{\theta}_{K-1},\boldsymbol{\theta}_K)$. Then as before $\prod_{i=1}^n g(Y_i,\mathbf{X}_i,\boldsymbol{\theta})$ can be maximized to find the ML estimate of $\boldsymbol{\theta}$. (Note that this can sometimes be computationally tricky.)

If any subset of the models in $(g_1(\boldsymbol{\theta}_1,\mathbf{X}_i),...,g_K(\boldsymbol{\theta}_1,\mathbf{X}_i))$ jointly characterize the conditional distribution of Y_i, then this estimator is consistent and asymptotically efficient. Although we do not expect this to actually be the case, it does imply that so long as we're close to the truth, we should have good performance.

Example 6.2.6. *Mixture Model Approximation*

Consider again the unusual distribution discussed in Example 6.2.4, visualized in Figure 6.2.3.

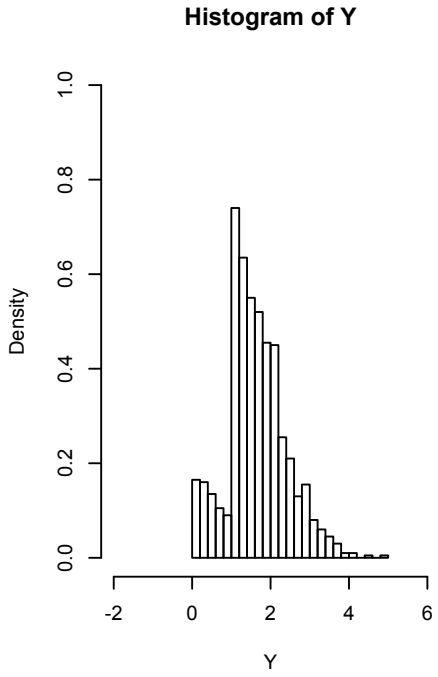


Figure 6.2.3 *Unusual Distribution*

In Example 6.2.4, we showed the consequences of modeling this distribution as normal:

$$g_N(y_i, (\mu_N, \sigma_N)) = \frac{1}{\sigma_N \sqrt{2\pi}} e^{-\frac{(y_i - \mu_N)^2}{2\sigma_N^2}}.$$

This model was obviously misspecified, but nevertheless a useful first-order approximation. But let us consider some other models to approximate this distribution. For example, we could model the distribution as log-normal:

$$g_{LN}(y_i, (\mu_{LN}, \sigma_{LN})) = \frac{1}{y_i \sigma_{LN} \sqrt{2\pi}} e^{-\frac{(\log y_i - \mu_{LN})^2}{2\sigma_{LN}^2}}.$$

Or we could model the distribution as exponential:

$$g_E(y_i, \lambda_E) = \lambda_E e^{-\lambda_E y_i}$$

The maximum likelihood estimates from each of these models are visualized in Figure 6.2.3.

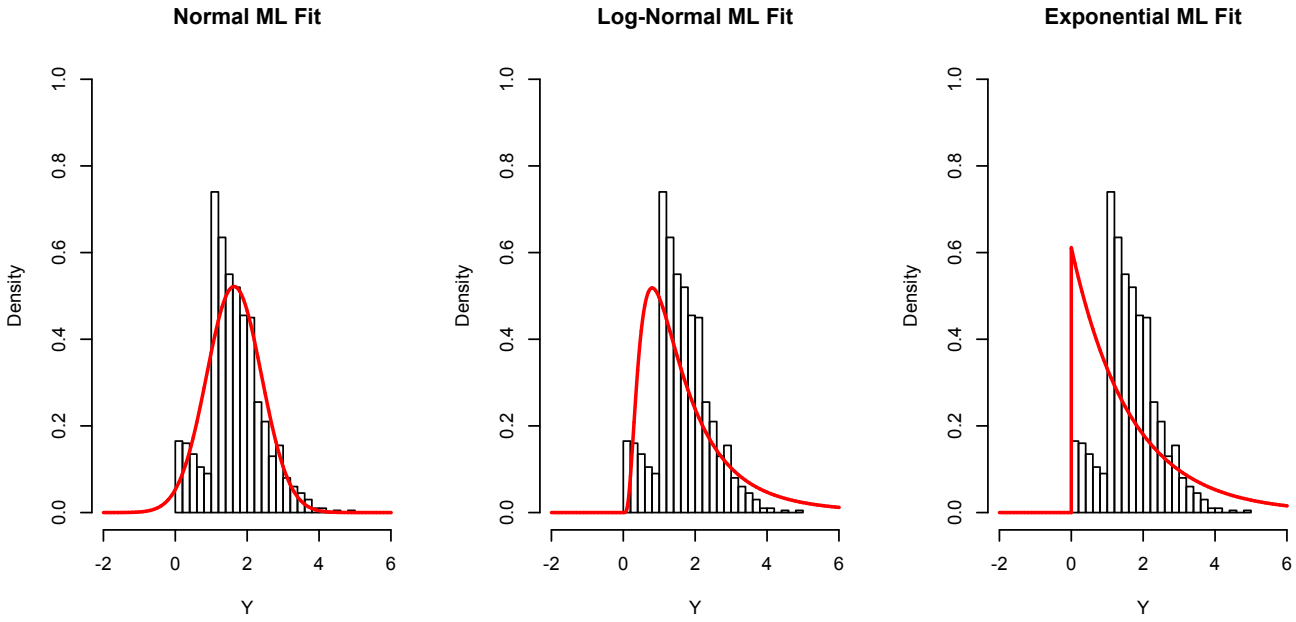


Figure 6.2.3 *Normal, Log-Normal, and Exponential Maximum Likelihood Approximations*

As we can see, all of these methods perform relatively poorly. But what happens when we combine them, and allow for the distribution of Y_i to be an optimal combination of the three distributions?

Consider a mixture model that incorporates all three models:

$$\begin{aligned} g_M(y_i, \boldsymbol{\theta}) &= p_1 g_N(y_i, (\mu_N, \sigma_N)) + p_2 g_{LN}(y_i, (\mu_{LN}, \sigma_{LN})) + (1 - p_1 - p_2) g_E(y_i, \lambda_E) \\ &= p_1 \frac{1}{\sigma_N \sqrt{2\pi}} e^{-\frac{(y_i - \mu_N)^2}{2\sigma_N^2}} + p_2 \frac{1}{y_i \sigma_{LN} \sqrt{2\pi}} e^{-\frac{(\log y_i - \mu_{LN})^2}{2\sigma_{LN}^2}} + (1 - p_1 - p_2) \lambda_E e^{-\lambda_E y_i}, \end{aligned}$$

where $\boldsymbol{\theta} = (p_1, p_2, \mu_N, \sigma_N, \mu_{LN}, \sigma_{LN}, \lambda_E)$. Figure 6.2.4 visualizes the maximum likelihood estimate of this mixture model, which provides a markedly better approximation.

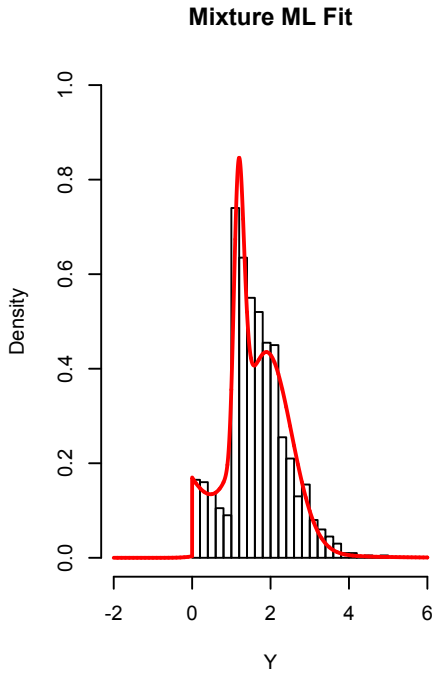


Figure 6.2.4 *Mixture Model Maximum Likelihood Approximation*

Technically, mixture models are parametric as the implied distribution of Y_i given $\mathbf{X}_i$ is fully governed by a finite number of parameters. But ML estimates of mixture models often have the operating characteristics of nonparametric estimators, and may often outperform explicitly nonparametric methods, including kernel methods.

In fact, depending on how we conceptualize their asymptotics, ML estimates of mixture models can be thought of as nonparametric estimators. If we imagined a sieve estimator where we added more and more models as n grew large in a sufficiently general way, then we would wind up having an infinite-dimensional model that could approximate any arbitrary distribution. Of course, once we have our ML estimate of the density function $f_{Y_i|\mathbf{X}_i}(y_i|\mathbf{x}_i)$, we again can do whatever we want with it, including characterizing the CEF, partial derivatives, or any other distributional feature.

6.2.7 Inference

A remaining question is how to characterize uncertainty for our ML estimates. There is debate over whether or not robust standard errors and/or the bootstrap are appropriate for parametric models. It is possible to estimate standard errors that exploit all of the parametric assumptions that we've used to generate the model. If we have estimates of the full distribution of Y_i conditional on X_i, we can use that information to construct a standard error estimate. This is the classical model-based approach. If the model is right, then we can directly use the model to estimate the standard errors, often with greater efficiency in finite samples.

The standard approach to model-based standard errors involves working off of a Taylor expansion of the log-likelihood function.

$$\hat{V}_{ML}(\hat{\boldsymbol{\theta}}) = \left[-\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} | \mathbf{Y}, \mathbb{X})}{\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\theta}}^T} \right]^{-1} \Bigg|_{\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}}.$$

The square root of the diagonals of the estimated variance-covariance matrix, $\hat{V}_{ML}(\hat{\boldsymbol{\theta}})$, are the associated standard errors. If the parametric model is correct, $\hat{V}_{ML}(\hat{\boldsymbol{\theta}})$ will consistently estimate the true variance of $\hat{\boldsymbol{\theta}}$.

Theorem 6.2.3. *Consistency of Model-Based Variance Estimation*

Given suitable regularity conditions (smoothness, compactness, unique maximum), if the parametric model is properly specified then

$$n\hat{V}_{ML}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} nV(\hat{\boldsymbol{\theta}}).$$

We omit the proof here, though a proof may be found in Newey and McFadden (2004).

There is an important caveat to this result: if the parametric model is incorrect, the model-based estimate $\hat{V}_{ML}(\hat{\boldsymbol{\theta}})$ will not generally be consistent for the true sampling variance. Model-based standard errors are not the only approach available for estimating standard errors of ML estimates, however. Robust standard errors are also possible, and they accurately characterize the sampling variability of $\hat{\boldsymbol{\theta}}$, even when the model is wrong. Furthermore, as usual, the bootstrap is available, and it too will yield consistent estimates of standard errors even when the model is wrong. Again, as usual, robust standard errors will asymptotically agree with the bootstrap, but the bootstrap will usually* have better finite n performance.

We know that the bootstrap and robust SEs will, with large n, give us the correct standard error for the ML estimate. The counter-argument to the use of the bootstrap or robust SEs is typically, "Who cares about getting the right standard error if your parametric model is wrong?" Well, we do! Parametric models are almost certainly wrong, but the estimates produced by ML estimation may still be meaningful as approximations. In Section 3.2.5, we considered "classical variance estimation" for the OLS estimator—an estimator that is logically equivalent to the model-based standard error for the ML estimate of the classical linear model. The estimates associated with OLS and other maximum likelihood solutions may have value as approximations even when the parametric model is incorrect. We should not needlessly run the risk of understating the sampling variability of these estimates. We believe that, given the near-certainty that any given parametric model is misspecified, for the types of problems that applied researchers face, robust standard errors or the bootstrap are typically preferable.

As long as you don't take the parametric model too seriously, ML estimation can be construed as a principled way of approximating a probability distribution. When we care about the features of the probability distribution (e.g., under a conditional independence assumption, some features of the distribution may be causal), then the ML estimates are best construed as good approximations to these features. Robust standard errors and the bootstrap guarantee, in large samples, that we are accurately quantifying the sampling variability of these approximations. Model-based standard errors offer no such guarantee.

6.3 Models as Approximations

We began this chapter with a famous quote by the statistician George Box. Many texts in statistics and econometrics include this quotation. In most cases, however, it is essentially taken to justify a more or less cavalier approach to statistical models, one that casually treats them as though they were true.

In this chapter, we have adhered to a more conservative interpretation of Box's aphorism. All models are wrong; they are, at best, approximations of reality. But, even without assuming that they are exactly true, when employed and interpreted correctly, they can nonetheless be useful for obtaining estimates of features of probability distributions.

7 Conclusion

We are all born naked, and the rest is drag.

— RUPAUL

What has been the purpose of this book? Clearly, we believe that the book provides a basis for statistical inference for the social and health sciences. But we hope it is something more than that. At its core, we have sought to reinforce a fundamental separation of statistical inference from the causal or structural assumptions that are used to relate statistical findings to real world behavior. In our view, statistical models are phenomenological approximations to observation, but only further assumptions can define their relationship to reality. Chapters 1, 2, 3, and 6 consider statistical inference; Chapters 4 and 5 consider classes of assumptions that allow us to link statistical inference to human behavior. These different tasks demand separate consideration; agnostic statistics finds its heart in this separation.

There is a language of statistics that we have introduced or reinforced for the reader. It is not the only language to describe the world, but it is an important one. Once we move past a bit of bothersome notation, it is not particularly complex.We believe this statistical literacy is essential for engaging in dialogues relating to empirical findings. These debates are often opaque and mired in jargon. But with a solid grounding in statistical theory, we have found that many of these conversations often reduce to simple propositions, often external to the statistical model.

We believe that this book will help to provide you with a basis for understanding how statistical methods work and how the researcher's assumptions shape the results. In recent years there has been a proliferation of new methods designed to solve substantive problems that social or health scientists face. Some of these methods may be useful in a given context; others may not. No statistical method is a panacea. These methods all embed both statistical assumptions and substantive assumptions, though sometimes the latter are not explicitly stated.

The ideas developed in this book should enable you to critically evaluate these methods by providing you with the conceptual tools to interpret their results under plausible assumptions. A solid grounding in agnostic statistics provides a basis for understanding the disjoint roles that statistical and causal assumptions play in constructing a body of empirical knowledge. When you read an empirical claim, you need not evaluate it solely in terms of the stated assumptions. Even findings that suffer from "endogeneity," selection bias, or any other failure of the stated assumptions may still convey useful descriptive information.

In terms of developing your own research program, we hope you approach questions with the humility demanded by agnosticism. Statistical inference is a means of description and quantification, and not on its own a basis for substantive knowledge. Our book is not a complete guide to empirical research, nor is it a document to be set in stone. Rather it is in many ways the beginning of a conversation that you the reader will have with the scientific community, your research subjects, and yourself.

As you proceed with your research, we encourage you to go forth and engage with the world, armed with an understanding of the role that statistical inference can assume when wedded to substantive assumptions. There is so much to be explored, to be learned, and to be shared. The practice of agnostic statistics, in its embrace of a fundamental uncertainty about the nature of the world, frees us to engage with these myriad

possibilities. We look forward to accompanying you on this journey of discovery, emancipated from the confines imposed by a restrictive model of the world.

A Glossary of Mathematical Notation

In this appendix, we provide a glossary for mathematical notation not otherwise defined in the main text.

Notation	Definition and Usage
$\in$	*Set membership.* $a \in A$ (read "a is an element of A" or "a is in A") denotes that the object a (which could be a number, a vector, a set, etc.) is an element of the set A. E.g., $2 \in \{1, 2, 4\}$. The negation of $\in$ is denoted by $\notin$. $b \notin A$ (read "b is not an element of A" or "b is not in A") denotes that the object b is *not* an element of the set A. E.g., $3 \notin \{1, 2, 4\}$.
$\forall$	*For all.* Used to state that all elements of a particular set satisfy a given expression. E.g., $\forall x \in \{1, 2, 3, 4, 5\}$, $2x \leq 10$. The set may be omitted if it is clear from context. E.g., $\forall x$, $x + 1 > x$ might mean that $x + 1 > x$ is true for any real number x. Similarly, $\forall k > 1, f(k) = 0$ might mean that $f(k) = 0$ is true for any integer k that's greater than 1. Usually, when defining a function, we will include a "for all" statement after the equation to denote the domain of the function. E.g., $g(x) = \sqrt{x}, \forall x \geq 0$.
$\exists$	*There exists.* Used to state that there is *at least one* element of a particular set satisfies a given expression. E.g., $\exists\, x \in \{1, 2, 3, 4, 5\}$ such that $x^2 = 16$. As with $\forall$, the set may be omitted if it is clear from context. E.g., $\exists\, x > 0$ such that $x^2 - 3x - 1 = 0$.
$\iff$	*If and only if.* Denotes that two statements are logically equivalent, i.e., each implies the other. Sometimes written as "iff." E.g., $2x = 10 \iff x = 5$.
$\subseteq$	*Subset.* $A \subseteq B$ (read "A is a subset of B") means that the set B contains every element in the set A. Formally, $A \subseteq B \iff \forall a \in A, a \in B$. E.g., $\{2, 3, 5\} \subseteq \{1, 2, 3, 4, 5\}$. Note that $A = B \iff A \subseteq B$ and $B \subseteq A$.
$\subset$	*Proper* (or *strict*) *subset.* $A \subset B$ (read "A is a strict subset of B") means that the set B contains every element in the set A *and* contains at least one element not in A. Formally, $A \subset B \iff \forall a \in A$, $a \in B$ and $\exists\, b \in B, b \notin A$.E.g., $\{1, 2, 4\} \subset \{1, 2, 3, 4, 5\}$.[104]
$\emptyset$	The *empty set*, i.e., the set containing no elements, sometimes written as $\{\}$.
$\setminus$	*Set subtraction.* $A \setminus B$ (read "A set-minus B") denotes the set that contains all the element in A *except* any that are also in B. Formally, $A \setminus B = \{a \in A : a \notin B\}$.
A^C	The *complement* of a set. When all sets under consideration are subsets of some *universal set* U, the complement of a set A is the set of all elements in U that are not in A. Formally, $A^C = U \setminus A$.

[104] Some texts use $\subset$ and $\subseteq$ interchangeably to denote the (non-strict) subset relation and use $\subsetneq$ to denote the strict subset relation. We prefer to treat $\subseteq$ and $\subset$ as analogous to $\leq$ and $<$, respectively. Likewise, $\supseteq$ and $\supset$ are analogous to $\geq$ and $>$.

$\cup$	The *union* of sets. $A \cup B$ (read "A union B") denotes the set containing all elements that are in *either* A *or* B (or both). Formally, $A \cup B = \{s : s \in A \text{ or } s \in B\}$. E.g., $\{1, 2, 5\} \cup \{2, 3\} = \{1, 2, 3, 5\}$. Union is associative, so for multiple unions parentheses can be omitted without ambiguity. E.g., $(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C$.
$\cap$	The *intersection* of sets. $A \cap B$ (read "A intersect B") denotes the set containing all elements that are in *both* A *and* B. Formally, $A \cap B = \{s : s \in A \text{ and } s \in B\}$. E.g., $\{1, 2, 5, 6\} \cap \{2, 3, 6\} = \{2, 6\}$. Intersection is associative, so for multiple intersections parentheses can be omitted without ambiguity. E.g., $(A \cap B) \cap C = A \cap (B \cap C) = A \cap B \cap C$.
$\mathcal{P}(A)$	The *power set* of A, i.e., the set of all subsets of A. E.g., $\mathcal{P}\Big(\{1, 2, 3\}\Big) = \Big\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\Big\}$.
$\lvert A \rvert$	The *cardinality* (or *size*) of a set. For finite sets, the cardinality is simply the number of elements in the set. E.g., $\lvert\{1, 2, 3, 5, 6, 8\}\rvert = 6$.
$\mathbb{N}$	The set of all *natural numbers*, i.e., positive integers: $\mathbb{N} = \{1, 2, 3, ...\}$. Note that while some texts include 0 in $\mathbb{N}$, we do not.
$\mathbb{Z}$	The set of all *integers*: $\mathbb{Z} = \{..., -3, -2, -1, 0, 1, 2, 3, ...\}$.[105]
$\mathbb{R}$	The set of all *real numbers*. In classical mathematics, there are many ways of rigorously defining the real numbers, but for our purposes it suffices to say that the real numbers are all the points on a continuous number line.
$\mathbb{R}^n$	The *real coordinate space* of dimension n, i.e., the set of all *vectors* of length n with real entries. Formally, $\mathbb{R}^n = \Big\{(x_1, x_2, ..., x_n) : x_i \in \mathbb{R}, \forall i \in \{1, 2, ..., n\}\Big\}$. $\mathbb{N}^n$, $\mathbb{Z}^n$, etc. are defined analogously.
$f : S \to T$	A *function* f from S to T. The set S is the *domain* of f, i.e., the set of all values s for which $f(s)$ is defined. The set T is the *codomain* of f, i.e., a set that contains all possible values of $f(s)$. Formally, $\forall s \in S, f(s) \in T$.
$f(A)$	The *image* of the set A under the function f, i.e., the set of all values the function f can take on when applied to an element of A. Formally, for a function $f : S \to T$ and a set $A \subseteq S$, $f(A) = \{t \in T : \exists\, a \in A \text{ such that } f(a) = t\}$. Note that $f(S)$ is the *range* of f.
$\sum$	*Summation* of a sequence. E.g., $\sum_{i=1}^{5} i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55$.
$\prod$	*Product* of a sequence. E.g., $\prod_{i=1}^{6}(i+1) = (1+1)(2+1)(3+1)(4+1)(5+1)(6+1) = 5040$.

[105] The letter 'Z' here stands for *Zahlen*, German for "numbers."

$\propto$	*Is proportional to.* $y \propto x$ (read "y is proportional to x") means that there is some nonzero constant k such that $y = kx$. Formally, $y \propto x \iff \exists k \in \mathbb{R} \backslash \{0\},\ y = kx$.
$\arg\max$	The value, among all values in a particular set, that minimizes the given expression. E.g., $\arg\max_{x \in \mathbb{R}} \left(-x^2 - 2x + 3\right) = -1$, as $-x^2 - 2x + 3$ attains its maximum value at $x = -1$.[106]
$\arg\min$	The value, among all values in a particular set, that minimizes the given expression. E.g., $\arg\min_{x \in \mathbb{R}} \left(x^2 - 4x + 1\right) = 2$, as $x^2 - 4x + 1$ attains its minimum value at $x = 2$.
$\mathbb{A}^T$	The *transpose* of a matrix. $\mathbb{A}^T$ denotes the matrix whose columns are the rows of $\mathbb{A}$. Sometimes written as $\mathbb{A}'$. E.g., if $\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$, then $\mathbb{A}^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$.
$\mathbb{A}^{-1}$	The *inverse* of a (square) matrix. If $\mathbb{A}$ is *invertible*, then $\mathbb{A}^{-1}$ denotes the matrix that, when multiplied by $\mathbb{A}$, yields the identity matrix of the appropriate dimensions: $\mathbb{A}\mathbb{A}^{-1} = \mathbb{A}^{-1}\mathbb{A} = \mathbb{I} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$ E.g., if $\mathbb{A} = \begin{pmatrix} 2 & 1 & 1 \\ -5 & -3 & 0 \\ 1 & 1 & -1 \end{pmatrix}$, then $\mathbb{A}^{-1} = \begin{pmatrix} -3 & -2 & -3 \\ 5 & 3 & 5 \\ 2 & 1 & 1 \end{pmatrix}$.

[106] Our use of $\arg\max$ ($\arg\min$) assumes there exists a unique maximum (minimum). More generally, $\arg\max$ ($\arg\min$) refers to the set of values at which the maximum (minimum) is attained.

B References

Abadie, Alberto and Guido W. Imbens. 2008. On the failure of the bootstrap for matching estimators. *Econometrica*. 76(6): 1537–1557.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Angrist, Joshua D. and Alan B. Krueger 1999. Empirical strategies in labor economics. *Handbook of Labor Economics*. 3: 1277–1366.

Aronow, Peter M. and Cyrus Samii. In press. Does regression produce representative estimates of causal effects? *American Journal of Political Science*.

Davidson, Russell and James G. MacKinnon. 2004. *Econometric theory and methods.*, Vol. 5. New York, NY: Oxford University Press.

Diaconis, Persi, Holmes, Sarah and Richard Montgomery. Dynamical bias in the coin toss. *Society for Industrial and Applied Mathematics Review*. 49(2): 211–235.

Efron, Bradley and Robert J. Tibshirani. 1994. *An introduction to the bootstrap*. Boca Raton, FL: CRC Press LLC.

Freedman, David A., Pisani, Robert and Roger A. Purves. 1998. *Statistics*, 3rd ed. New York: Norton.

Freedman, David A. 2009. *Statistical models: theory and practice*. New York, NY: Cambridge University Press.

Goldberger, Arthur S. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Hájek, Jaroslav. 1971. Comment on 'An essay on the logical foundations of survey sampling, Part I.' In Godambe, Vidyadhar P. and David A. Sprott (Eds.). *Foundations of Statistical Inference*. Toronto: Holt, Rinehart, and Winston.

Hansen, Bruce E. 2013. *Econometrics*. Manuscript, University of Wisconsin.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association*. 81(396): 945–968.

Horvitz, Daniel G. and Donovan J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 47: 663–684.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press.

Newey, Whitney. K. and Daniel McFadden. 1994. Chapter 36: Large sample estimation and hypothesis testing. *Handbook of Econometrics*. 4: 2111–2245.

Jaynes, Edwin T. 2003. *Probability theory: The logic of science*. London: Cambridge University Press.

Kang, Joseph D. Y. and Joseph L. Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating population means from incomplete data. *Statistical Science*. 22(4): 523–539.

Lovell, Michael C. 2008. A Simple Proof of the FWL Theorem. *Journal of Economic Education*. 39(1): 88–91.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. New York, NY: Springer-Verlag New York, Inc.

Morgan, Stephen L. and Christopher Winship. 2014. *Counterfactuals and Causal Inference*. New York, NY: Cambridge University Press.

Splawa-Neyman, Jerzy, Dabrowska, Dorota M. and Terence P. Speed. 1923. Reprint, 1990. On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statistical Science*. 5(4): 465–472.

Robins, James M. and Andrea Rotnitzky. 2001. Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer.' *Statistica Sinica*. 11(4): 920–936.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 70(1): 41–55.

Silverman, Bernard W. 1986. *Density estimation for statistics and data analysis*. London; New York, NY: CRC Press.

Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 58(1): 267–288.

Wasserman, Larry. 2004. *All of statistics: A concise course in statistical inference*. New York, NY: Springer Science+Business Media, Inc.

Wasserman, Larry. 2006. *All of nonparametric statistics*. New York, NY: Springer Science+Business Media, Inc.

Wasserman, Larry. 2012. Lecture notes for 10-705/36-705 Intermediate Statistics.