BUSN 41901 / STAT 32400

Probability and Statistics

Lecture Notes

Autumn 2023

Tetsuya Kaji

University of Chicago Booth School of Business

Contents

Preface

This is a collection of lecture notes for the first-year graduate course in probability and statistics for economics and business students. Major emphasis is put on asymptotic (large-sample) statistics. The intended reader is a user of statistics rather than a developer, so proofs are kept minimal and more focus is placed on the interpretation and connection of theory and applications.

Theorem clauses contain the main theoretical exposition of the course. Propositions are addenda that are outside the scope of the exams. Remarks contain miscellaneous notes—minor or too advanced—that may be disregarded without affecting the main argument. Chapter appendices are also outside the scope of the exams.

There are some notational conventions we employ in the lecture notes. The indicator function is denoted by $\mathbb{1}\{\cdot\}$, for example, $\mathbb{1}\{x = 5\}$ equals 1 if $x = 5$ and 0 otherwise. The maximum and minimum of two numbers are denoted with $\vee$ and $\wedge$, so $a \vee b \coloneqq \max\{a, b\}$ and $a \wedge b \coloneqq \min\{a, b\}$. The norm $\|\cdot\|$ denotes the Euclidean norm. Unless otherwise specified, a vector is a column vector. The ordering of two vectors is elementwise, i.e., $x \geq y$ means $x_i \geq y_i$ for every index i. Symmetric positive semidefinite matrices are ordered by the *Loewner ordering*, that is, $A \geq B$ means that $A - B$ is positive semidefinite. The prime applied to a vector gives a transpose; to a function gives a derivative. For example, for a matrix-valued function $f : \mathbb{R} \to \mathbb{R}^{n\times m}$, $f'(t)'$ denotes the transpose of the $n \times m$ matrix of derivatives $f'(t) = \frac{d}{dt} f(t)$. Double differentiation with respect to a vector argument is understood as a Hessian, i.e., for a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$, $f''(x) = \frac{d^2}{dxdx'} f(x)$ is an $n \times n$ matrix of second-order cross derivatives. When a function depends on two arguments θ and x, one of them may be occasionally put as a subscript, e.g., $\ell_\theta(x)$. The partial differentiation of ℓ with respect to the subscript argument is denoted with a dot, i.e., $\dot{\ell}_\theta(x) = \frac{\partial}{\partial\theta}\ell_\theta(x)$ and $\ddot{\ell}_\theta(x) = \frac{\partial^2}{\partial\theta\partial\theta'}\ell_\theta(x)$. The Kronecker product is denoted by $\otimes$, e.g.,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{bmatrix}.$$

For a square matrix A, $\det(A)$ denotes the determinant of A and $\operatorname{tr}(A)$ the trace of A. For a matrix A, $\operatorname{vec}(A)$ denotes the vectorization of A into a column vector, e.g., $\operatorname{vec}\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = (a, c, b, d)'$. For a symmetric matrix A, $\operatorname{vech}(A)$ denotes the half-vectorization of its lower triangular part, e.g., $\operatorname{vech}\left(\begin{smallmatrix} a & b \\ b & d \end{smallmatrix}\right) = (a, b, d)'$. The operator diag

converts vectors and square matrices into diagonal matrices, e.g., $\mathrm{diag}(a, b) = \left[\begin{smallmatrix} a & \\ & b \end{smallmatrix}\right]$ and $\mathrm{diag}\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left[\begin{smallmatrix} a & \\ & d \end{smallmatrix}\right]$.

A limited number of applied papers are summarized and presented throughout the notes to illustrate the concepts. Simplicity is prioritized over accuracy. For detailed and precise description of what they do, I encourage the reader to dig into the original papers.

CHAPTER 1

Introduction

Believe those who are seeking the truth; doubt those who find it; doubt everything, but don't doubt of yourself.

André Gide, translated by justin o'brien, 1952

1.1. Why Statistics?

Why do we learn statistics? The simplest answer is because statistics is the basis of inductive reasoning, and inductive reasoning is key to today's social science.

Let us review the overarching structure of science. American philosopher Charles Sanders Peirce classified reasoning into three categories: deduction, induction, and abduction. *Deduction* is the process of extracting consequences that follow from accepted premises. If A implies B and B implies C, then concluding that A implies C is a deductive reasoning. Mathematics is an obvious example of the use of deduction.[1] The merit of deductive reasoning is its rigor. *Induction* is the process of drawing general conclusions from examples. With the observation that the sun has risen from the east thus far, concluding that the sun will rise from the east tomorrow is an inductive reasoning. An example in science is when a medical trial finds the efficacy of a new drug. A benefit of inductive reasoning is that it allows us to be agnostic on the rigorous chain of steps by which one thing leads to another. *Abduction* is the process of forming explanatory hypotheses. If A implies B and B is observed, then proposing A as a possible explanation for B is an abductive reasoning. A prominent example is when a physicist comes up with a new theory of everything. Abductive reasoning enables us to generate new insights and new perspectives.

One mode of reasoning does not stand alone. A theory abductively proposed in physics must be followed by a load of experiments to verify its implications, whose process is mostly inductive. A drug inductively discovered by a medical experiment invites further research to find out why it is effective, which may be abductive.

Economics is no exception to utilizing various inferential arguments; there are subfields of economics whose principal modes of inquiry vary across all three. A prominent subfield of a deductive nature is microeconomic theory; that of abductive is behavioral economics. However, being the science of as complex a matter as humans

[1] The method of *mathematical induction* is yet a form of deduction.

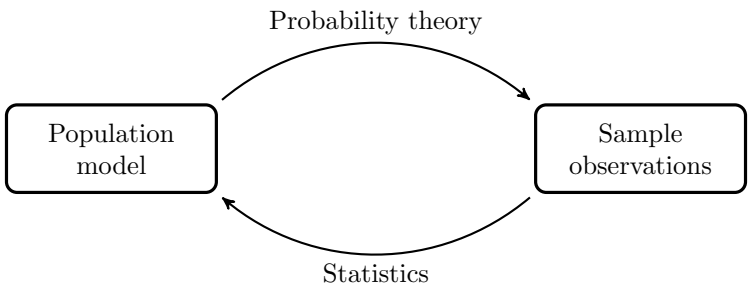


FIGURE 1.1. Probability theory and statistics.

and their behaviors, a substantial portion of the economics discipline is dedicated to the realm of *applied research*, whose principal mode of inquiry is induction. It is no surprise that the major advantage of inductive reasoning as being independent of the underlying mechanism plays a crucial role in the credibility of economic research. Familiarizing yourself with the basic knowledge of statistics, therefore, is an integral and imperative part of graduate education in economics and business.

1.2. What is Statistics?

The starting point of inductive reasoning is to collect data, and their observation is often made in a probabilistic manner. Income may be determined by a number of factors such as intelligence and social skills, which may seem random for the factors unobservable to us; customer survey may only be feasible on a subset of customers and the subset may be chosen randomly. In this sense, probability and statistics are closely connected, even the flip sides of the same coin (Figure 1.1).

Probability theory concerns how random events take place, how data come about from a more fundamental probabilistic structure. However, what we are interested in is usually the fundamental structure and not the data themselves; as concerned as we are about the underlying relationship of the wage and education, the specific wage and education level of a specific individual are of no interest to economists. *Statistics*, in that regard, aims to uncover the fundamental structure from the data. In this sense, statistics can be considered an inverse operation of probability realization. In the early days of statistics, the field was indeed called "inverse probability."[2]

Questions that statistics can answer can be classified into three types.

1. *Descriptive:* What is the prevalence of the flu? What is the compliance rate of transactions in a company? What errors do peer reviewers detect?
2. *Predictive:* What will the trajectory of a Hurricane be? What will be the revenue of a company next quarter? Which movie will a customer like?
3. *Causal:* Does reducing the class size improve elementary school education? Is there racial discrimination in the market for home loans? How much do cigarette taxes reduce smoking?

[2]Statistics started with what is now known as Bayesian inference. So "inverse probability" is considered a precursor to Bayesian statistics. Yet, the conceptualization of inverting the direction of the arrow stands just as well for frequentist statistics.

Contemporary academic research in economics and business is heavily geared toward causal inference, while statistics in industry is equally focused on prediction.

1.3. Which Statistics?

Not only that there is uncertainty in the observation of data—which is formulated as probability—there can just as well be uncertainty in the probabilistic structure itself. For example, while the outcome of a coin toss is random, there is almost no uncertainty that it follows a Bernoulli distribution. But when it comes to the firm size of an American firm, what distribution it follows is very much disputable. If we end up making a wrong assumption about the distribution, the statistical analysis that follows might be inaccurate.

In light of this, it is important to distinguish the two types of statistics: finite-sample statistics and asymptotic (large-sample) statistics. *Finite-sample statistics* builds on a complete description of the probability structure, and the characteristics of our statistical analyses can be driven precisely. In this sense, finite-sample statistics is exact. It is useful when the statistician can decide on the entire randomness in small experiments or when the sample size is small as in psychology. On the other hand, *asymptotic statistics* uses an approximation to the probability distributions in exchange for weaker assumptions. Such is possible when many distinct probability structures give rise to a similar situation as the size of the data gets bigger. It is useful for observational studies or large experiments in which a vast number of data points are available but specific distributional assumptions are eschewed, as in economics and business research.

This course puts more emphasis on asymptotic statistics as it is the widely accepted viewpoint of applied research in economics and business. Note, however, that finite-sample statistics and asymptotic statistics are not mutually exclusive. When one statistical method can be analyzed by either, the assumptions required by them are often nested. In that case, statistical modeling can be made in two layers of validity—when strong assumptions hold, we can draw exact conclusions in finite samples, and if we only buy into weak assumptions, we can still draw approximate conclusions whose precision depends on how large the dataset is.

CHAPTER 2

Probability Theory

But to us, probability is the very guide of life.

JOSEPH BUTLER, QUOTED IN THE ANALOGY OF RELIGION, INTRODUCTION BY HENRY G. BOHN, 1852

Among all kinds of random objects, real-valued random variables are the basics of all and most important. In this chapter, we discuss some properties of real-valued univariate and multivariate random variables.

2.1. Probability Distributions

2.1.1. Univariate distributions. The probability distribution of a one-dimensional random variable is described by the cumulative distribution function.

Definition 2.1 (Cumulative distribution function)**.** The *cumulative distribution function (cdf)* of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) \coloneqq P(X \leq x).$$

The cdf gives a complete characterization of any univariate random variable.

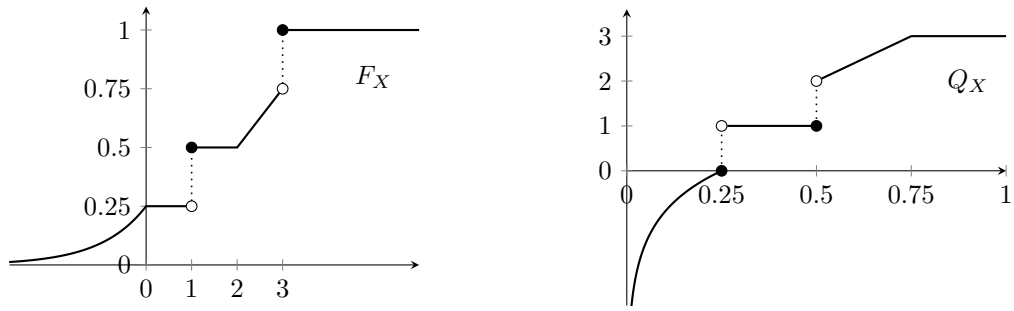
Theorem 2.1 ([CB02, Theorem 1.5.3])**.** *A function $F : \mathbb{R} \to [0, 1]$ is a cdf of some random variable if and only if F is nondecreasing and right-continuous and satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.*

The inverse of the cdf is called the quantile function. The quantile function is defined on an open interval $(0, 1)$ to avoid singularities at the end points. This gives a percentile for any percentage normalized between 0 and 1. It is conventionally defined as a *left*-continuous function, unlike the cdf (Figure 2.1).

Definition 2.2 (Quantile function)**.** The *quantile function* of a random variable X is the left-continuous generalized inverse $Q_X : (0, 1) \to \mathbb{R}$ of F_X,

$$Q_X(u) \coloneqq \inf\{x \in \mathbb{R} : F_X(x) \geq u\}.$$

EXAMPLE 2.1 (Bernoulli). A random variable X is called *Bernoulli* if $X = 1$ with some probability p and $X = 0$ with probability $1-p$. We have $F_X(x) = (1-p)\mathbb{1}\{x \geq 0\} + p\mathbb{1}\{x \geq 1\}$ and $Q_X(u) = \mathbb{1}\{u > 1-p\}$. A coin toss is an example of a Bernoulli random variable with $p = 1/2$.



(A) Probability distribution function. (B) Quantile function.

FIGURE 2.1. Probability distribution function and quantile function of a random variable X that is neither continuous nor discrete.

EXAMPLE 2.2 (Uniform). A random variable X is called *uniform* if $F_X(x) = 0 \vee \frac{x-a}{b-a} \wedge 1$ for some $a < b$. We denote it by $X \sim U[a, b]$. The quantile function of X is $Q_X(u) = u$. A lottery is an example of a uniform random variable.

The cdf may not be the most convenient tool to work on a random variable with. Alternatively, we can define a random variable by the derivative of a cdf. If such a representation exists, we call the random variable absolutely continuous. In statistics, however, absolutely continuous random variables are casually called continuous.

Definition 2.3 (Discrete and continuous random variables)**.** The random variable X is said to be *discrete* if F_X is a step function. It is said to be *continuous* if F_X is continuous. It is said to be *(absolutely) continuous* if F_X is absolutely continuous, that is, there exists a function $p_X : \mathbb{R} \to \mathbb{R}$ such that for every $x \in \mathbb{R}$,

$$F_X(x) = \int_{-\infty}^{x} p_X(t)dt.$$

In this case, p_X is called the *probability density function (pdf)* of X.

Theorem 2.2 ([CB02, Theorem 1.6.5])**.** *A function $p : \mathbb{R} \to \mathbb{R}$ is a pdf of some random variable if and only if $p \geq 0$ and $\int_{-\infty}^{\infty} p(t)dt = 1$.*

EXAMPLE 2.3 (Normal). A random variable X is called *normal* if the pdf takes the form $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$. We denote it by $X \sim N(\mu, \sigma^2)$. The cdf and quantile function of X do not have closed-form expressions. For example, the height of a randomly chosen person from a homogeneous population can be very well approximated by a normal random variable.

Remark 2.1. It is possible to extend the notion of pdf to any random variable, absolutely continuous or not, by the means of a Radon–Nikodym derivative (Section 2.A). In fact, the "probability mass function (pmf)" that you may have learned in other statistics courses is a "pdf" with respect to some non-Lebesgue measure.

2.1.2. Multivariate distributions. When there is more than one random variable, their relationship must be defined by the means of a *joint* distribution.

Definition 2.4 (Multivariate cdf and pdf)**.** The *joint cdf* of k real-valued random variables $X_1, \ldots, X_k$ is $F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) := P(X_1 \le x_1, \ldots, X_k \le x_k)$. The *joint pdf* of $X_1, \ldots, X_k$, if exists, is a function $p_{X_1,\ldots,X_k} : \mathbb{R}^k \to \mathbb{R}$ such that

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} p_{X_1,\ldots,X_k}(s_1, \ldots, s_k) ds_k \cdots ds_1.$$

The univariate cdf F_X and pdf p_X are sometimes specifically called the *marginal cdf* and *marginal pdf* if there is a need for distinction. By definition, we can calculate the marginal cdf from the joint cdf, e.g., as

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x, y).$$

If a joint pdf exists, then we can calculate the marginal cdf and pdf by

$$F_X(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} p_{X,Y}(s, t) dt ds, \qquad p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, t) dt.$$

The most important relationship between random variables is that there is no relationship, which is called independence.

Definition 2.5 (Independence)**.** The random variables $X_1, \ldots, X_k$ are called *independent* if their joint cdf is represented in the product form, $F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = F_{X_1}(x_1) \cdots F_{X_k}(x_k)$. If such decomposition exists, each function F_{X_j} corresponds to the marginal cdf of X_j. Denote by $X_1 \perp X_2$ if X_1 and X_2 are independent.

Exercise 2.1 (Pairwise independence). Let (X, Y, Z) be a triplet of binary random variables that takes four values $\{(0, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 0)\}$ with equal probability. Show that (X, Y) are independent, (Y, Z) are independent, and (Z, X) are independent, but (X, Y, Z) are not independent. It is said that (X, Y, Z) are *pairwise* independent but not *mutually* (or *jointly*) independent.

If the joint cdf admits a joint pdf, then $F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = F_{X_1}(x_1) \cdots F_{X_k}(x_k)$ if and only if $p_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = p_{X_1}(x_1) \cdots p_{X_k}(x_k)$.

We sometimes employ the vector notation. For example, if $X = (X_1, \ldots, X_k)'$ is a $k \times 1$ column vector of random variables, then $F_X(x) = P(X_1 \le x_1, \ldots, X_k \le x_k)$ denotes the joint cdf, where $x = (x_1, \ldots, x_k)'$ is also a $k \times 1$ vector.

Remark 2.2. For a k-dimensional random vector $X = (X_1, \ldots, X_k)'$, the function $C(u_1, \ldots, u_k) := F_X(F_{X_1}^{-1}(u_1), \ldots, F_{X_k}^{-1}(u_k))$ is called the *copula* and has some application in finance.

2.2. Functions of Random Variables

Many statistical applications involve transformations of random variables.

Example 2.4 (Sealed-bid auction). Each bidder of an auction has an associated value, monetary or utility, that she gains by winning the auction. The function that maps the value to the bid is called the bidding strategy. From an economist's

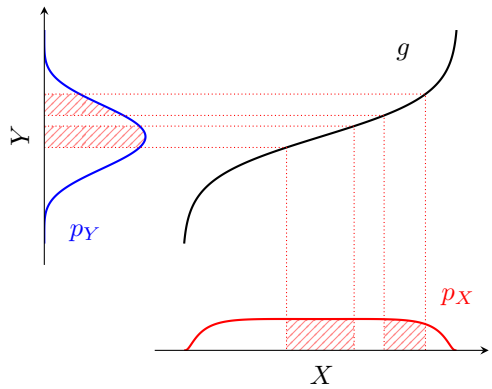


FIGURE 2.2. The change of variables $Y = g(X)$. The steeper g leads to the greater p_Y; the flatter g to the less p_Y.

perspective, the randomness of the bids comes from the randomness of the values, and it is often of interest to figure out the distribution of the values from the distribution of the bids, which, e.g., allows counterfactual analysis of alternative auction systems.

EXAMPLE 2.5 (Option pricing). The payoff of a financial derivative is given as a transformation of the payoff of the underlying asset. For example, the payoff of a European call option of a stock with a strike price k is given by $\max\{X - k, 0\}$ where X is the price of the underlying stock at the expiration date. To price the option, the distribution of the payoff must be derived from the distribution of X.

For a k-dimensional random variable X and a suitably measurable function $g : \mathbb{R}^k \to \mathbb{R}^m$, $Y := g(X)$ defines a new m-dimensional random variable. Its cdf is

$$F_Y(y) = P(g(X) \leq y) = P(X \in g^{-1}((-\infty, y_1] \times \cdots \times (-\infty, y_m])).$$

If $k = m = 1$, g is increasing and differentiable at x, and F_X is differentiable at x, the pdf of Y at $y = g(x)$ can be computed by the chain rule

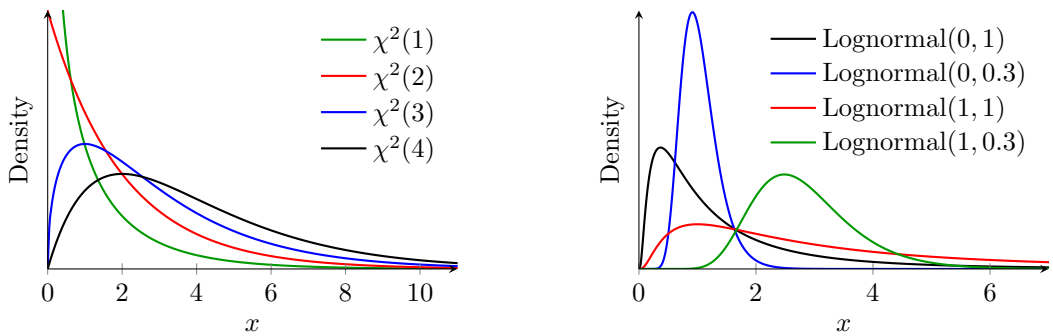
$$p_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(g^{-1}(y))}{dy} = p_X(g^{-1}(y))\frac{1}{g'(g^{-1}(y))}.$$

Figure 2.2 provides the intuition that the steeper the slope of g, the smaller the transformed density. This can be extended to cases where $k = m > 1$ and g is bijective but not necessarily increasing as

$$p_Y(y) = p_X(g^{-1}(y))\left|\det\left(\frac{dg^{-1}(y)}{dy'}\right)\right|.$$

EXAMPLE 2.6 (Inverse transform sampling). Let $U \sim U[0, 1]$ and Q_X be a quantile function. Then the cdf of $X := Q_X(U)$ is given by $F_X = Q_X^{-1}$. This is one way computers generate random variables from various distributions.

EXERCISE 2.2 (Probability integral transform). Let X be a continuous random variable and F_X its cdf. Show that $U := F_X(X)$ follows $U[0, 1]$. This is related to the distribution of p-values (Exercise 5.4).



(A) Chi-square pdfs. (B) Lognormal pdfs.

FIGURE 2.3. Transformations of the normal distribution.

EXAMPLE 2.7 (Linear transformation and normality). Let $X = (X_1, \ldots, X_k)' \sim N(\mu, \Sigma)$ be a multivariate normal vector. Then, any linear combination of X follows a univariate normal distribution, that is, for every $\beta \in \mathbb{R}^k$, $X'\beta$ follows $N(\mu'\beta, \beta'\Sigma\beta)$. The converse is also true: if $X'\beta$ follows a univariate normal distribution for every $\beta \in \mathbb{R}^k$, then X follows a multivariate normal distribution. This property can in turn be used to "extend" the notion of normal distribution to more complicated random objects, such as Banach-valued normal distributions.

EXERCISE 2.3 (Convolution). Let

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right)$$

be a bivariate normal vector with $-1 < \rho < 1$. Derive the pdf of $Z := X + Y$. *Hint: Derive the joint pdf of* $\left[\begin{smallmatrix} X \\ Z \end{smallmatrix}\right] = \left[\begin{smallmatrix} X \\ X+Y \end{smallmatrix}\right]$ *using the above formula and take the marginal of* Z.

EXAMPLE 2.8 (Chi-square distribution). If $X_1, X_2, \ldots, X_k$ are independent standard normal random variables, the distribution of $Y = X_1^2 + \cdots + X_k^2$ is known as the *chi-square distribution with* k *degrees of freedom* and is denoted by $\chi^2(k)$ (Figure 2.3A). For two independent χ^2 random variables Y_1 and Y_2 with degrees of freedom k_1 and k_2, the distribution of the ratio $\frac{Y_1/k_1}{Y_2/k_2}$ is known as the F*-distribution with* k_1 *and* k_2 *degrees of freedom.*

EXAMPLE 2.9 (Lognormal distribution). If X follows a normal distribution $N(\mu, \sigma^2)$, the distribution of $\exp(X)$ is known as the *lognormal distribution* and is written as $\text{Lognormal}(\mu, \sigma^2)$ (Figure 2.3B). The right tail of a lognormal distribution is known to decay slower than an exponential function.

Independence is preserved under transformations.

Theorem 2.3 (Preservation of independence)**.** *If* X *and* Y *are independent, then for every measurable function* g, $g(X)$ *and* Y *are independent.*

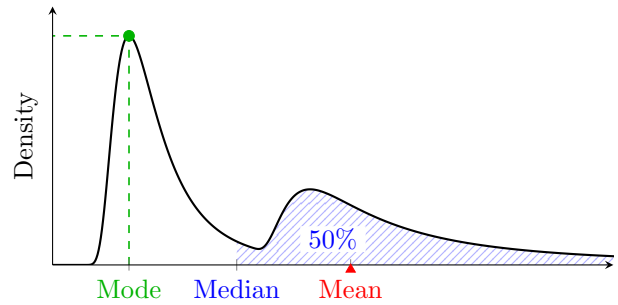


FIGURE 2.4. Mean is where the scale of pdf balances; median is where the probability is split in half; mode is where the pdf attains its maximum.

2.3. Expectation and Moments

When we face probabilistic situations, it is often helpful to look at some quantities that summarize specific aspects of the randomness. For example, a marketer may evaluate promotional offers by the expected revenue they bring, or an investor may pick a portfolio based on the expected return and the risk of various stocks. Many of these summary quantities are defined by the means of expectation.

Definition 2.6 (Expectation)**.** For a random variable X in $\mathbb{R}^k$ and a suitably measurable function $g : \mathbb{R}^k \to \mathbb{R}^\ell$, the *expectation* of $g(X)$ is given by

$$\mathbb{E}[g(X)] := \int_{\mathbb{R}^k} g(x) dF_X(x).$$

Theorem 2.4 (Law of the unconscious statistician)**.** *Let* $Y := g(X)$*. Then* $\mathbb{E}[Y] = \mathbb{E}[g(X)]$*.*

The expectation of a random variable is a primary measure of its location. However, alternative location measures do exist and are sometimes more useful (Figure 2.4). For example, the median income is used to describe the income of an average individual; a tail quantile is used to measure the risk of a portfolio (Value-at-Risk).

Definition 2.7 (Median and quantiles)**.** The *median* $\mathrm{Med}(X)$ of a univariate random variable X is $Q_X(1/2)$. For $\alpha \in (0, 1)$, the *αth quantile* of X is $Q_X(\alpha)$.

EXERCISE 2.4. Show that $\mathbb{E}[X] = \int_0^1 Q_X(u)du$.

Expectations of some powers of a random variable are called moments.

Definition 2.8 (Moments)**.** For a positive integer r, the *rth moment* of a univariate random variable X is defined by

$$\mathbb{E}[X^r] := \int_{\mathbb{R}} x^r dF_X(x),$$

and the *rth central moment* of X is defined by $\mathbb{E}[(X - \mathbb{E}[X])^r]$. For a real number r, the *rth absolute moment* of X is defined by $\mathbb{E}[|X|^r]$.

EXERCISE 2.5. Prove that if $\mathbb{E}[|X|^r] < \infty$ for some $r > 0$, then $\mathbb{E}[|X|^q] < \infty$ for every $0 < q < r$.

The second central moment of a random variable measures the dispersion and is known as variance.

Definition 2.9 (Variance)**.** The *variance* of a univariate random variable X is

$$\mathrm{Var}(X) \coloneqq \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{\mathbb{R}} (x - \mathbb{E}[X])^2 dF_X(x).$$

The *standard deviation* of X is $\sqrt{\mathrm{Var}\,(X)}$.

EXAMPLE 2.10 (Tail probability). The standard deviation can be used to bound the tail probability of a random variable. Let $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}(X)$. If X is normally distributed, the *three-sigma rule* tells that $P(|X - \mu| \geq \sigma) \approx 32\%$, $P(|X-\mu| \geq 2\sigma) \approx 5\%$, and $P(|X-\mu| \geq 3\sigma) \approx 0.3\%$. If X is not normally distributed, this relationship no longer holds. However, we can still (very loosely) bound the tail probability by the variance through Chebyshev's inequality, for $c > 0$,

$$P(|X - \mu| \geq c\sigma) \leq \frac{\mathbb{E}[(X - \mu)^2]}{c^2\sigma^2} = \frac{1}{c^2}.$$

EXERCISE 2.6. For each $c \geq 1$, find a random variable X such that $P(|X - \mu| \geq c\sigma) = 1/c^2$. This means that Chebyshev's inequality cannot be improved without additional assumptions on X.

Variance is by no means the only measure of dispersion. For example, the expected absolute deviation from the mean, $\mathbb{E}[|X - \mathbb{E}[X]|]$, also measures how widespread the distribution of X is. So why is variance our favorite dispersion measure? Here are some pros and cons.

Mean squared deviation (variance):

- 👍 Relates nicely to the mean (Theorem 2.5). This gives justification to the use of a conditional mean as the best predictor (Section 7.1.2).
- 👍 Relates nicely to the dispersion of the sum of random variables. For example, $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$ if $X_1, \ldots, X_n$ are independent.
- 👍 Relates nicely to the central limit theorem (Theorem 3.5). An average inherits variance from each component, but not other dispersion measures.
- 👍 Relates nicely to the concept of covariance (Definition 2.10). This is surprising given that covariance is a measure of *comovement*, which sounds conceptually irrelevant to a measure of *dispersion*.
- 👍 Can exploit orthogonality. We can make use of the mathematical wisdom such as the Pythagorean theorem and orthogonal projection.
- 👍 Compatible with a large body of statistics.
- 👎 Hard to interpret. If X is in the units of \$, the variance is in the units of $\$^2$, but what does that mean?

Mean absolute deviation (MAD):

- 👍 Can exist even when variance does not. This gives a kind of robustness in some applications.
- 👍 Relates nicely to the concept of the median (Theorem 2.6).
- 👍 Relates nicely to some financial concepts (straddle strategy, etc.).
- 👎 Compatible with a small body of statistics.

Overall, the variance wins and hence the primary dispersion measure in statistics.

A notable relationship of the mean and variance is the minimum-minimizer relationship given as follows.

Theorem 2.5 (Mean minimizes squared error)**.** *If* $\mathbb{E}[X^2] < \infty$*, then we have* $\mathbb{E}[X] = \arg\min_{b\in\mathbb{R}} \mathbb{E}[(X-b)^2]$ *and* $\mathrm{Var}(X) = \min_{b\in\mathbb{R}} \mathbb{E}[(X-b)^2]$.

Proof. Write $\mathbb{E}[(X-b)^2] = \mathbb{E}[(X-\mathbb{E}[X]+\mathbb{E}[X]-b)^2] = \mathbb{E}[(X-\mathbb{E}[X])^2]+2\mathbb{E}[(X-\mathbb{E}[X])(\mathbb{E}[X]-b)] + \mathbb{E}[(\mathbb{E}[X]-b)^2]$. The first term is irrelevant to minimization. The second term is zero. The third term is uniquely minimized at $b = \mathbb{E}[X]$. ■

An analogous relationship holds between the median and the mean absolute deviation.

Theorem 2.6 (Median minimizes absolute error)**.** *If* $\mathbb{E}[|X|] < \infty$*, then we have* $\mathrm{Med}(X) \in \arg\min_{b\in\mathbb{R}} \mathbb{E}[|X-b|]$*. More generally, if* $\mathbb{E}[|X|] < \infty$*, then* $Q_X(\tau) \in \arg\min_{b\in\mathbb{R}} \mathbb{E}[\rho_\tau(X-b)]$ *where* $\tau \in (0,1)$ *and* $\rho_\tau(u) := u(\tau - \mathbb{1}\{u<0\})$ *is the* check function *for the* τ*th quantile.*

Exercise 2.7. Prove the first statement of Theorem 2.6 for the case where X is absolutely continuous with positive density. *Hint: Use the Leibniz integral rule.*

The relationship of multiple random variables can also be quantified using expectation. The following gives a measure of comovement between two random variables.

Definition 2.10 (Covariance)**.** The *covariance* of univariate random variables X and Y is

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \int_{\mathbb{R}^2} (x-\mathbb{E}[X])(y-\mathbb{E}[Y])dF_{X,Y}(x,y).$$

The *correlation* of X and Y is $\mathrm{Corr}(X,Y) := \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}\,(X)\,\mathrm{Var}(Y)}$.

Covariance is large when X and Y tend to deviate to the same direction (positive or negative) from the corresponding means, but also when the scales of X and Y are large. For example, if you want to measure the correlation of prices of an Uber ride and a Lyft ride, the covariance becomes 10,000 times larger by switching the units from dollars to cents. Correlation removes this scale dependence by dividing by the standard deviations. It is immediate by the Cauchy–Schwarz inequality that $-1 \le \mathrm{Corr}(X,Y) \le 1$.

Exercise 2.8. Variance and covariance relate nicely with each other. For instance, the variance of a sum $X+Y$ can be calculated with the knowledge of the variances and covariance of X and Y. Show that $\mathrm{Var}(aX+bY) = a^2\,\mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X,Y) + b^2\,\mathrm{Var}(Y)$.

It is straightforward to generalize the variance and covariance to vectors.

Definition 2.11 (Variance and covariance for vectors)**.** The *variance* of a $k \times 1$ random vector X is a $k \times k$ matrix

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])'].$$

The *covariance* of k- and ℓ-dimensional random vectors X and Y is a $k \times \ell$ matrix

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

For a vector X, the diagonal elements of $\mathrm{Var}(X)$ are the variances of the individual components of X, and the off-diagonal elements are the covariances of the pairs of the components. For this, $\mathrm{Var}(X)$ is sometimes called the variance-covariance matrix.

There are alternative formulas for the variance and covariance, which are simpler to deal with in some cases.

Theorem 2.7 (Properties of variance)**.** *For random vectors X and Y with finite variances, the following hold.*

(i) $\mathrm{Var}(X) = \mathbb{E}[XX'] - \mathbb{E}[X]\mathbb{E}[X]'$.
(ii) $\mathrm{Cov}(X, Y) = \mathbb{E}[XY'] - \mathbb{E}[X]\mathbb{E}[Y]'$.
(iii) If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Exercise 2.9. Prove Theorem 2.7.

Remark 2.3. The expectation of the product $\mathbb{E}[XY']$ is called the *cross moment* of X and Y.

Exercise 2.10. For a bivariate normal random variable (X, Y), show that X and Y are independent if and only if they are uncorrelated.

Exercise 2.11. Construct a pair of random variables (X, Y) such that each of them is marginally standard normal, they are uncorrelated, but they are not independent.

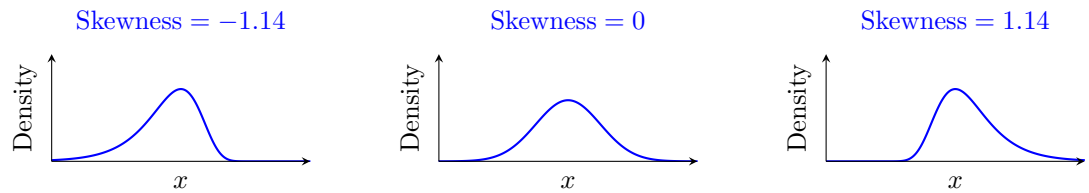
The central scaled third and fourth moments are known as the skewness and kurtosis and appear in finance.

Definition 2.12 (Skewness and kurtosis)**.** The *skewness* of a univariate random variable X is given by $\mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}\right)^3\right]$. The *kurtosis* of X is given by $\mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}\right)^4\right]$.

The skewness is positive when the distribution is pinched to the right and negative when pinched to the left (Figure 2.5). The kurtosis is large when the distribution has heavy tails and small when it has light tails (Figure 2.6).

2.4. Conditional Expectation and Conditional Probability

It is often the case that we want to infer a random variable with the knowledge of another. For example, a labor economist may be interested in how the wage is determined by education and experience (Mincer equation); a financial economist may want to know whether the announcement of a merger leads to abnormal stock returns (event study) or if low interest rates induce risk-taking behaviors ("reaching for yield").

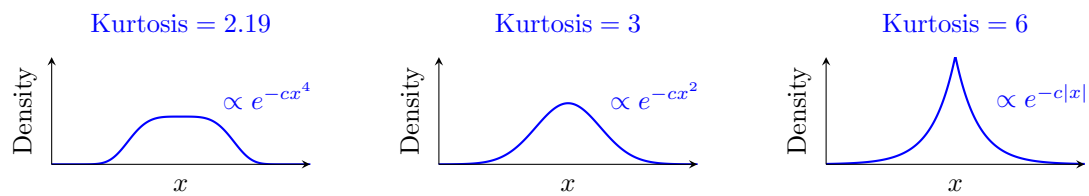


(A) Minimum Gumbel distribution.

(B) Standard normal distribution.

(C) Maximum Gumbel distribution.

FIGURE 2.5. Skewness measures asymmetry.



(A) Quartic exponential power distribution.

(B) Standard normal distribution.

(C) Laplace distribution.

FIGURE 2.6. Kurtosis measures the tail thickness.

Updating of the randomness due to the knowledge of another is precisely described by the notion of conditional probability and conditional expectation. When the conditioning event has a positive probability, the classical ratio formula $P(A \mid B) = P(A \cap B)/P(B)$ gives the conditional probability of event A conditional on event B. However, we want to have a more general definition that does not require a positive denominator. Standard construction of such generalization builds on measure theory and is quite complicated. Here, I give a version that would be more accessible. Interestingly, the general definition of conditional expectation precedes that of conditional probability.

Definition 2.13 (Conditional expectation)**.** For k- and ℓ-dimensional random vectors X and Y and a measurable function $g : \mathbb{R}^k \to \mathbb{R}^\ell$, $g(X)$ is called the *conditional expectation of* Y *given* X if $\mathbb{E}[(Y - g(X))'h(X)] = 0$ for every measurable bounded function $h : \mathbb{R}^k \to \mathbb{R}^\ell$. In this case, we denote $g(X)$ by $\mathbb{E}[Y \mid X]$ and $g(x)$ by $\mathbb{E}[Y \mid X = x]$.

Remark 2.4. $\mathbb{E}[Y \mid X = x]$ is well defined almost everywhere even if $P(X = x) = 0$ for every x.

This can readily be extended to more involved conditioning; for example, $\mathbb{E}[Y \mid X \geq 0]$ can be defined as $\mathbb{E}[Y \mid Z = 1]$ with a new random variable $Z \coloneqq \mathbb{1}\{X \geq 0\}$.

EXERCISE 2.12. Let

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right)$$

be a bivariate normal vector with $-1 < \rho < 1$. Show that $\mathbb{E}[Y \mid X = x] = \rho\frac{\sigma_Y}{\sigma_X}x$. *Hint: We want to show that* $\mathbb{E}[(Y - \rho\frac{\sigma_Y}{\sigma_X}X)h(X)] = 0$ *for every bounded* h. *Derive the joint pdf of* $\begin{bmatrix} X \\ Y-\rho\frac{\sigma_Y}{\sigma_X}X \end{bmatrix}$ *as in Exercise 2.3 and verify that* $Y - \rho\frac{\sigma_Y}{\sigma_X}X$ *is mean zero and independent of* X.

Exercise 2.13. Show that $\mathbb{E}[f(X)Y \mid X = x] = f(x)\mathbb{E}[Y \mid X = x]$. You may assume for simplicity that f is real-valued, bounded, and strictly positive.

Theorem 2.8 (Existence of conditional expectation)**.** *Let* X, Y *be random variables. If* $\mathbb{E}[|Y|] < \infty$, *then* $\mathbb{E}[Y \mid X]$ *exists.*

Remark 2.5. The converse does not hold. Let Y be Cauchy and $X = Y$. Then $\mathbb{E}[Y \mid X] = X$ and $\mathbb{E}[|Y|] = \infty$.

The following says that the expectation of conditional expectation is expectation.

Theorem 2.9 (Law of iterated expectations)**.** *Let* X *and* Y *be random variables. If* $\mathbb{E}[|Y|] < \infty$, *then* $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$.

Proof. Let $h \equiv 1$ in the definition of conditional expectation. ■

Exercise 2.14. Denote $g(x) = \mathbb{E}[Y \mid X = x]$ and $h(y) = \mathbb{E}[X \mid Y = y]$. Assuming that g is injective, show that if $h = g^{-1}$, then X and Y are perfectly dependent, that is, $Y = g(X)$.

There is an explicit formula for the conditional expectation for absolutely continuous variables.

Theorem 2.10 (Conditional expectation formula)**.** *Let* X *and* Y *be univariate random variables and* $\mathbb{E}[|Y|] < \infty$. *If* (X, Y) *has a joint pdf* $p_{X,Y}$, *then*

$$\mathbb{E}[Y \mid X = x] = \frac{\int_{-\infty}^{\infty} t p_{X,Y}(x,t)dt}{\int_{-\infty}^{\infty} p_{X,Y}(x,t)dt} = \frac{\int_{-\infty}^{\infty} t p_{X,Y}(x,t)dt}{p_X(x)},$$

provided that the denominator is not zero.

Proof. Let h be bounded and denote $\int_{-\infty}^{\infty}$ and $p_{X,Y}$ by $\int$ and p. By Fubini's theorem,

$$\mathbb{E}\left[\left(Y - \frac{\int tp(X,t)dt}{\int p(X,t)dt}\right)h(X)\right] = \iint yh(x)p(x,y)dxdy - \int \frac{\int tp(x,t)dt}{\int p(x,t)dt}h(x)\int p(x,y)dydx = 0.$$

■

It is known that the best predictor of Y given X in terms of the mean squared deviation is the conditional expectation of Y given X. Theorem 2.5 is a corollary of this when X is constant.

Theorem 2.11 (Projection)**.** *If* $\mathbb{E}[Y^2] < \infty$, *then* $\mathbb{E}[Y \mid X = x] = \arg\min_{g(\cdot)} \mathbb{E}[(Y - g(X))^2]$ *where* g *runs through all measurable functions.*

PROOF. Let g be bounded.[1] Write $\mathbb{E}[(Y - g(X))^2] = \mathbb{E}[(Y - \mathbb{E}[Y \mid X] - g(X) + \mathbb{E}[Y \mid X])^2] = \mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2] - 2\mathbb{E}[(Y - \mathbb{E}[Y \mid X])(g(X) - \mathbb{E}[Y \mid X])] + \mathbb{E}[(g(X) - \mathbb{E}[Y \mid X])^2]$. The first term is irrelevant for minimization. The second term is zero by the definition of conditional expectation. The third term is minimized at $g(X) = \mathbb{E}[Y \mid X]$. ■

The notion of conditional probability is defined through the conditional expectation of an indicator.

Definition 2.14 (Conditional probability)**.** Let X and Y be k- and ℓ-dimensional random vectors. The *conditional probability of an event A given $X = x$* is defined by $P(A \mid X = x) \coloneqq \mathbb{E}[\mathbb{1}\{A\} \mid X = x]$. Specifically, the *conditional cdf of Y given $X = x$* is defined by $F_{Y|X}(y \mid x) \coloneqq \mathbb{E}[\mathbb{1}\{Y \leq y\} \mid X = x]$. If $F_{Y|X}(\cdot \mid x)$ is absolutely continuous, the *conditional pdf of Y given $X = x$* is defined analogously to Definition 2.4 and denoted by $p_{Y|X}(\cdot \mid x)$.

EXERCISE 2.15. Show that if X and Y are independent, the conditional cdf of Y given X is the same as the marginal cdf of Y.

An explicit formula for the conditional pdf for absolutely continuous variables is available.

Theorem 2.12 (Conditional density formula)**.** *Let X and Y be univariate random variables. If (X, Y) has a joint pdf $p_{X,Y}$, then*

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x, y)}{\int_{-\infty}^{\infty} p_{X,Y}(x, t)dt} = \frac{p_{X,Y}(x, y)}{p_X(x)},$$

provided that the denominator is not zero.

PROOF. Applying Theorem 2.10 to $F_{Y|X}(y \mid x) = \mathbb{E}[\mathbb{1}\{Y \leq y\} \mid X = x]$, we find

$$F_{Y|X}(y \mid x) = \frac{\int_{-\infty}^{\infty} \mathbb{1}\{t \leq y\} p_{X,Y}(x, t)dt}{p_X(x)} = \frac{\int_{-\infty}^{y} p_{X,Y}(x, t)dt}{p_X(x)}.$$

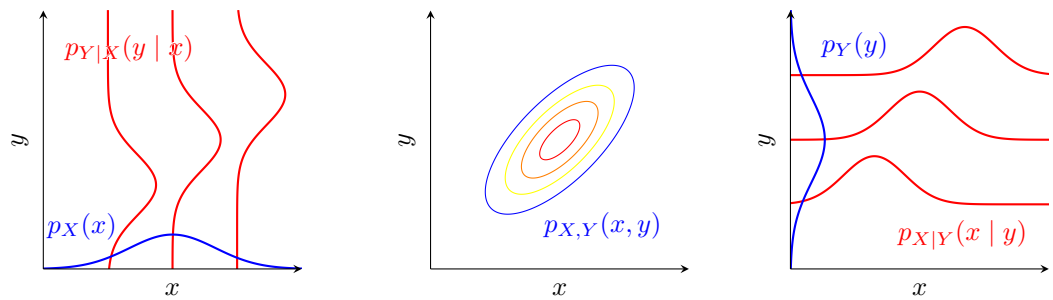
Then, the claim follows by the Leibniz integral rule. ■

This formula indicates that there are three equivalent ways to represent the joint distribution (Figure 2.7),

$$p_X(x)p_{Y|X}(y \mid x) = p_{X,Y}(x, y) = p_Y(y)p_{X|Y}(x \mid y).$$

The left formula sees the randomness of (X, Y) in sequence; we first observe X, and then observe Y with the knowledge of X. This representation is suitable when there is a direction $X \to Y$, e.g., when we observe the characteristics of a used car (X) and want to predict its price (Y), or when we raise the corporate tax (X) and want to know its consequences on the corporate behavior (Y). The middle formula sees their randomness on simultaneous, equal terms. This is suitable when there is no prespecified direction of observation, e.g., when we want to optimize the portfolio over various assets (X and Y). The right formula is the same as the left with the

[1] Since bounded functions are dense in L^1, this is without loss of generality.



(A) $p_X(x)p_{Y|X}(y \mid x)$. This decomposition emphasizes how the knowledge of X affects the uncertainty of Y.

(B) $p_{X,Y}(x, y)$. This representation emphasizes the joint uncertainty of X and Y.

(C) $p_Y(y)p_{X|Y}(x \mid y)$. This decomposition emphasizes how the knowledge of Y affects the uncertainty of X.

FIGURE 2.7. Three equivalent representations of the distribution of X and Y.

roles of X and Y reversed. Note that the left representation does *not* require that X *realize* before Y; it is just that X be *observed* before Y.

Another implication of this is Bayes's rule: $p_{Y|X}(y \mid x) = p_Y(y)p_{X|Y}(x \mid y)/p_X(x)$. This says that the uncertainty of Y given X can be recovered if we know the uncertainty of X given Y and their marginals. For example, when we have a macroeconomic model $p_{X|Y}(x \mid y)$ of how economic parameters Y affect economic variables X and have a prior $p_Y(y)$ on the parameters Y, we can derive the distribution $p_{Y|X}(y \mid x)$ of the parameters that are consistent with the observed data X [**HS16**].[2]

EXERCISE 2.16. When X is continuous and Y is discrete, create a figure as in Figure 2.7 to illustrate the three representations.

Conditional variance is defined analogously to conditional expectation.

Definition 2.15 (Conditional variance)**.** For random vectors X and Y, the *conditional variance of* Y *given* X is defined by $\mathrm{Var}(Y \mid X) := \mathbb{E}[(Y - \mathbb{E}[Y \mid X])(Y - \mathbb{E}[Y \mid X])' \mid X]$ if exists.

The following is a decomposition of the uncertainty $\mathrm{Var}(Y)$ into the uncertainty of prediction $\mathrm{Var}(\mathbb{E}[Y \mid X])$ and the prediction of the uncertainty after prediction $\mathbb{E}[\mathrm{Var}(Y \mid X)]$.

Theorem 2.13 (Law of total variance)**.** *Let* X *and* Y *be random vectors such that* $\mathbb{E}[YY'] < \infty$. *Then* $\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y \mid X]) + \mathbb{E}[\mathrm{Var}(Y \mid X)]$.

EXERCISE 2.17. Prove Theorem 2.13.

[2]The marginal distribution $p_X(x)$ of X is only for scaling and actually not needed.

2.A. Defining Densities for Arbitrary Random Variables

The density is one of the most fundamental concepts in probability theory, yet its definition requires a nontrivial adjustment for random variables that are not absolutely continuous.

Consider a Bernoulli random variable X that takes value 1 with probability p and 0 with $1-p$. The cdf of X is $F_X(x) = (1-p)\mathbb{1}\{x \le 0\} + p\mathbb{1}\{x \le 1\}$, which is discontinuous and not differentiable. Therefore, the expectation of $f(X)$ is given as a sum, not as a familiar integral, i.e.,

$$\int_{\mathbb{R}} f(x)dF_X(x) = (1-p)f(0) + pf(1).$$

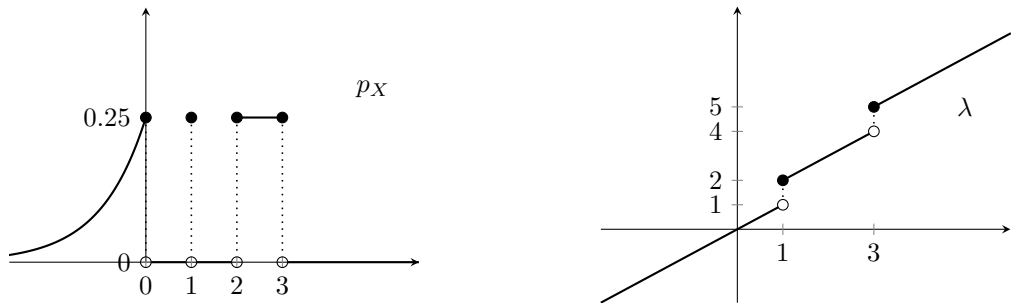
However, we can still define the density of X if we use a non-Lebesgue measure. Let μ be the measure that has unit masses at 0 and 1, so its "distribution function" is $\mu(x) = \mathbb{1}\{x \le 0\} + \mathbb{1}\{x \le 1\}$. Then, for a function $p_X(x) = p^x(1-p)^{1-x}$, we have

$$\int_{\mathbb{R}} f(x)dF_X(x) = \int_{\mathbb{R}} f(x)p_X(x)d\mu,$$

so p_X acts as the "pdf" of X with respect to μ. This μ is called the *dominating measure* and p_X the *Radon–Nikodym derivative* of the distribution of X with respect to μ. Of course, this p_X is not the only function that satisfies this property. For example, we could have very well defined it as $(1-p)\mathbb{1}\{x = 0\} + p\mathbb{1}\{x = 1\}$ or $(1-p) + (2p-1)x$ and enjoyed the same density property. This non-uniqueness is usually not an issue, so we can pick one that is easier to handle. For example, the first definition is nice when we consider the logarithm of the density.

Note that this also depends on the choice of the dominating measure. If we use a measure λ that has the Lebesgue measure plus unit masses at 0 and 1 (so its integral is given by $\int_{\mathbb{R}} f(x)d\lambda = \int_{-\infty}^{\infty} f(x)dx + f(0) + f(1)$), then $p_X(x) = (1-p)\mathbb{1}\{x = 0\} + p\mathbb{1}\{x = 1\}$ is the only function (among the three discussed) that satisfies the density property. Non-uniqueness of the dominating measure is neither a problem; rather, paramount is the fact that for any two random variables—be it continuous, discrete, or whatever—we can always find a common dominating measure such that both random variables have densities with respect to it.

A virtue of the pdf is that it represents the likeliness of X at one single point, as opposed to the cdf representing the likeliness of realizing at anywhere below a point. So, for a given realized value x, we can now compare the "likelihood" of any two probability distributions, take the ratio, or even differentiate them. Figure 2.8 shows one choice of the dominating measure and the density for the random variable defined in Figure 2.1.



(A) A density of X with respect to λ.

(B) Distribution function of λ.

FIGURE 2.8. The random variable X in Figure 2.1 is not absolutely continuous but has a density with respect to the measure λ in (B).

CHAPTER 3

Asymptotic Theory of Probability

Nothing is more uncertain *than the duration of* individual *life: nothing is more* certain *than the* average *continuance of life.*

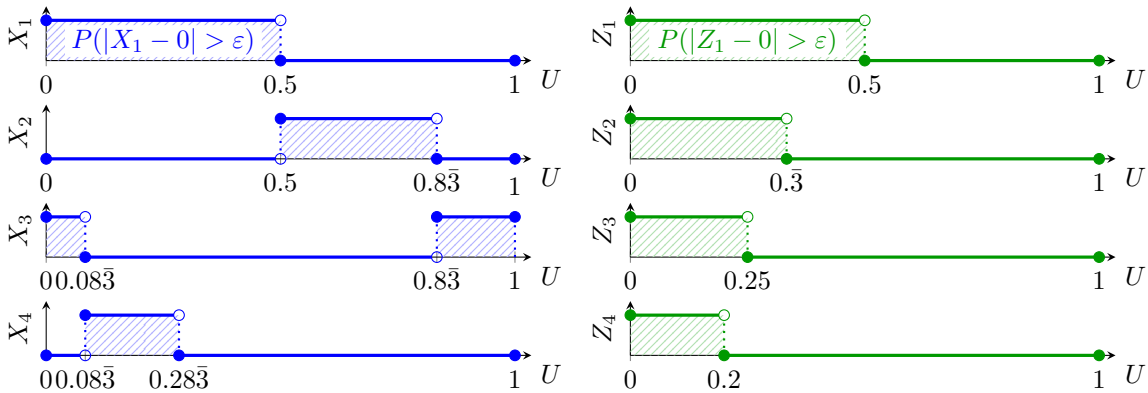
INSURANCE GUIDE AND HANDBOOK, CORNELIUS WALFORD, 1868

Asymptotic statistics gives justification to many statistical methods used in social science; it shows that various reasonable but distinct probability structures lead to a common simplified structure when the sample size is large. This observation legitimizes the use of the simplified structure as an approximation while remaining open to various possibilities of different probability structures.

As such, the limit theorems in this section are stated in terms of the limit as the sample size "n" goes to infinity. This does not mean that the actual sample size of the data must continue to grow. A useful analogy is the linear approximation of a smooth function. Let us say that a smooth function admits a linear approximation at a point x_0. We know that the precision of this approximation improves as the point of evaluation gets closer and closer to the point of expansion x_0. However, this does not mean that the linear approximation is only useful *at* x_0. (Indeed, there is no need to approximate it at x_0.) Instead, the linear approximation is useful *around* x_0 but we just need to keep in mind that the precision of the approximation depends on how close the point of evaluation is to x_0. Similarly, despite the limit theorems being stated as $n \to \infty$, the approximation still stands for a finite and non-growing dataset.

3.1. Modes of Convergence

Unlike the convergence of a deterministic sequence, the convergence of a random sequence takes place in many different ways. There are three modes of convergence that are usually introduced in a graduate level statistics course—almost sure convergence, convergence in probability, and convergence in distribution. For an educated user of statistics, it is important to understand the latter two; for a developer of statistics, the first one ise sometimes used in proofs. For completion, I introduce all three.



(A) X_n in Example 3.1 converges to 0 in probability but not almost surely. For every $u \in [0,1]$, $X_n(u)$ fluctuates between 0 and 1 infinitely many times as $n \to \infty$, albeit less and less frequently.

(B) Z_n in Exercise 3.1 converges to 0 in probability and almost surely. In statistical applications, it is practically impossible to tell apart X_n from Z_n in many cases.

FIGURE 3.1. Almost sure convergence and convergence in probability.

Definition 3.1 (Almost sure convergence)**.** A sequence of random vectors X_n *converges almost surely* to a random vector X, written $X_n \to^{\text{as}} X$, if

$$P\Big(\lim_{n\to\infty} \|X_n - X\| = 0\Big) = 1.$$

Definition 3.2 (Convergence in probability)**.** A sequence of random vectors X_n *converges in probability* to a random vector X, written $X_n \to^p X$, if for every $\varepsilon > 0$

$$\lim_{n\to\infty} P(\|X_n - X\| > \varepsilon) = 0.$$

Remark 3.1. $X_n \to^p X$ is also denoted as $\operatorname{plim}_{n\to\infty} X_n = X$.

Remark 3.2. Almost sure convergence can equivalently be formulated as

$$P\Big(\lim_{n\to\infty} \|X_n - X\| > \varepsilon\Big) = 0$$

for every $\varepsilon > 0$. This reveals that the only difference between $\to^{\text{as}}$ and $\to^p$ is the order of lim and P.

Almost sure convergence implies convergence in probability as we will see below (Theorem 3.2 (i)), but not the converse.

EXAMPLE 3.1 (Counterexample to almost sure convergence). Let $U \sim U[0,1]$ and define $X_n(U) = \mathbb{1}\{\sum_{k=1}^{n} \frac{1}{k} \le U + m < \sum_{k=1}^{n+1} \frac{1}{k}, \exists m \in \mathbb{N}\}$ (Figure 3.1A). Then, for every realization $u \in [0,1]$ of U, $X_n(u)$ hits 1 infinitely many times, although less and less often. This means that $X_n(u)$ does not converge for every u and hence the probability that $\lim_{n\to\infty} |X_n(U) - 0| = 0$ takes place is zero, that is, $X_n \not\to^{\text{as}} 0$. On the other hand, for every small $\varepsilon > 0$ and fixed n, we have $P(|X_n(U) - 0| > \varepsilon) = \frac{1}{n+1}$, which converges to 0. Therefore, $X_n \to^p 0$.

The following exercise explains why almost sure convergence is of less interest.

EXERCISE 3.1 (Almost sure representation). Let $U \sim U[0,1]$ and define $Z_n(U) = \mathbb{1}\{U < \frac{1}{n+1}\}$ (Figure 3.1B). Show that Z_n has the same marginal distribution as X_n in Example 3.1 and $Z_n \to^{\text{as}} 0$. In statistical applications, the "sample size" n is fixed, and there is no way to tell apart X_n from Z_n by observation.

The last mode of convergence is the most important.

Definition 3.3 (Convergence in distribution)**.** A sequence of random vectors X_n *converges in distribution* to a random vector X, written $X_n \rightsquigarrow X$, if for every real-valued bounded continuous function f,

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

Remark 3.3. By analogy, we also write $X_n \rightsquigarrow F$ to mean that $X_n \rightsquigarrow X$ for $X \sim F$. Convergence in distribution is also called *weak convergence* or *convergence in law* and may be denoted with $\to^d$ or $\Rightarrow$. I personally prefer $\rightsquigarrow$ because it not only looks "weak" but nicely combines the "distributed as" symbol $\sim$ with the convergence arrow $\to$.

There are many equivalent formulations of the convergence in distribution. While Definition 3.3 is readily generalizable to various random objects such as random functions, the following alternative is often handier when it comes to real-valued random vectors.

Theorem 3.1 (Portmanteau lemma)**.** *Let X_n and X be random vectors. Then, $X_n \rightsquigarrow X$ if and only if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for every x at which F_X is continuous.*

PROOF. [**vdV98**, Lemma 2.2]. ■

EXAMPLE 3.2 (Fraud counts). Let p be the probability that a credit card transaction is fraudulent. For n transactions, let X_n be the number of frauds. Then, X_n follows a binomial distribution with parameters n and p, i.e., $P(X_n = k) = {}_nC_k p^k (1-p)^{n-k}$. Suppose that the fraud detection algorithm improves at the same time the customer base grows, so $p = 1/n$ as $n \to \infty$. Then, X_n converges in distribution to a Poisson distribution with parameter 1. To see this, note that

$$P(X_n = k) = \frac{n!}{(n-k)!k!} n^{-k} \left(1 - \frac{1}{n}\right)^{n-k}.$$

Using Stirling's approximation,

$$\frac{n!}{(n-k)!} n^{-k} \approx \left(\frac{n}{n-k}\right)^{1/2} e^{-k} \left(1 + \frac{k}{n-k}\right)^{n-k} \longrightarrow 1.$$

Thus, we find $P(X_n = k) \to e^{-1}/k!$ and the convergence follows from Theorem 3.1.

EXERCISE 3.2. Consider a sequence of random variables X_n whose cdf is given by $F_{X_n}(x) = 1 - (1 - x/n)^n$ for $0 \le x \le n$. Find F such that $X_n \rightsquigarrow F$.

EXERCISE 3.3. Consider a sequence of random variables X_n whose pdf is given by $1 - \cos(n\pi x)$ for $x \in [0,1]$. Show that the pdf of X_n does not converge but X_n converges in distribution.

The three modes of convergence are nested as follows.

Theorem 3.2 (Relation of modes of convergence)**.** *Let c be a nonrandom constant.*

(i) $X_n \to^{\text{as}} X$ *implies* $X_n \to^p X$.
(ii) $X_n \to^p X$ *implies* $X_n \rightsquigarrow X$.
(iii) $X_n \rightsquigarrow c$ *implies* $X_n \to^p c$.
(iv) $X_n \rightsquigarrow X$ *and* $Y_n \rightsquigarrow c$ *imply* $X_n + Y_n \rightsquigarrow X + c$ *and* $X_n Y_n \rightsquigarrow cX$.

Proof. (i) For every n, we have $\sup_{m \geq n} P(\|X_m - X\| > \varepsilon) \leq P(\sup_{m \geq n} \|X_m - X\| > \varepsilon)$. Thus, $\limsup_{n\to\infty} P(\|X_n - X\| > \varepsilon) \leq P(\limsup_{n\to\infty} \|X_n - X\| > \varepsilon)$.

(ii) I prove the univariate case. See [**vdV98**, Theorem 2.7 (ii)] for a general proof. For every $\varepsilon > 0$,

$$\begin{aligned} F_{X_n}(x) &= P(X_n \leq x, X \leq x + \varepsilon) + P(X_n \leq x, X > x + \varepsilon) \\ &\leq P(X \leq x + \varepsilon) + P(|X_n - X| > \varepsilon) = F_X(x + \varepsilon) + P(|X_n - X| > \varepsilon). \end{aligned}$$

Similarly, $F_X(x-\varepsilon) \leq F_{X_n}(x) + P(|X_n - X| > \varepsilon)$. Thus, $F_X(x-\varepsilon) - P(|X_n - X| > \varepsilon) \leq F_{X_n}(x) \leq F_X(x+\varepsilon) + P(|X_n - X| > \varepsilon)$. Since $X_n \to^p X$, we have $P(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$. Therefore, $\limsup_n F_{X_n}(x)$ and $\liminf_n F_{X_n}(x)$ are bounded by $F_X(x - \varepsilon)$ and $F_X(x + \varepsilon)$ for every $\varepsilon > 0$. If x is a continuity point of F_X, then $F_{X_n}(x) \to F_X(x)$ by the squeeze theorem.

(iii) I prove the univariate case. See [**vdV98**, Theorem 2.7 (iii)] for a general proof. Since every $x \neq c$ is a continuity point of F_c, if $X_n \rightsquigarrow c$, we have $F_{X_n}(x) \to \mathbb{1}\{x > c\}$. Then, $P(|X_n - c| > \varepsilon) \leq P(X_n \leq c - \varepsilon) + P(X_n > c + \varepsilon) = F_{X_n}(c - \varepsilon) + 1 - F_{X_n}(c + \varepsilon)$, which converges to 0 for every $\varepsilon > 0$.

(iv) [**vdV98**, Lemma 2.8]. ■

Remark 3.4. Theorem 3.2 (iv) is known as *Slutsky's lemma*.

Another useful fact is that each mode of convergence is preserved under continuous transformations.

Theorem 3.3 (Continuous mapping theorem; [**vdV98**, Theorem 2.3])**.** *Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at almost every realization of X. Then, the following hold.*

(i) $X_n \to^{\text{as}} X$ *implies* $g(X_n) \to^{\text{as}} g(X)$.
(ii) $X_n \to^p X$ *implies* $g(X_n) \to^p g(X)$.
(iii) $X_n \rightsquigarrow X$ *implies* $g(X_n) \rightsquigarrow g(X)$.

Exercise 3.4. Convergence in distribution does not imply convergence of moments [**Dav21**, Section 23.4]. Construct a sequence X_n that converges in distribution to a standard normal distribution but $\mathbb{E}[X_n]$ and $\mathbb{E}[X_n^2]$ do not converge to 0 and 1.

3.1.1. Stochastic big O small o notation. In Taylor expansion, the lower-order terms start to dominate the higher-order terms as the point of evaluation approaches the point of expansion, that is, the linear term dominates the quadratic, and the quadratic dominates the cubic. In other words, the quadratic term is more ignorable than the linear term, and the cubic term is more ignorable than the linear

and quadratic terms. Similar situations arise in statistics. In statistics, the "approaching the point of expansion" corresponds to the sample size going to infinity. This is because when the sample size is infinity, it is as if the true population model is observed, and hence it makes a reasonable evaluation point for a large dataset.

In doing so, some approximation errors are more ignorable than others, and the following big O small o notation becomes handy. Intuitively, $X_n = o_P(R_n)$ means that X_n is more ignorable than R_n, and $X_n = O_P(R_n)$ means that X_n is as ignorable (or as important) as R_n. The precise definition is as follows.

Notation. The notation $X_n = o_P(1)$ means that $X_n \to^p 0$. The notation $X_n = o_P(R_n)$ for a sequence of (possibly random) variables R_n means that $X_n = Y_n R_n$ for some $Y_n = o_P(1)$. The notation $X_n = O_P(1)$ means that X_n is *uniformly tight*, that is, for every $\varepsilon > 0$ there exists M such that $\sup_n P(\|X_n\| > M) < \varepsilon$. The notation $X_n = O_P(R_n)$ means that $X_n = Y_n R_n$ for some $Y_n = O_P(1)$.

Example 3.3. For $X_n \sim U[0, 1/n]$, we have $X_n = o_P(1)$ and $X_n = O_P(1/n)$. For $X_n \sim U[n, n+1]$, we have $X_n = O_P(n)$, $X_n - n = O_P(1)$, and $X_n = o_P(n^2)$.

Exercise 3.5. Show that $X_n \to^p X$ implies $X_n - X = o_P(1)$ and that $X_n \rightsquigarrow X$ implies $X_n = O_P(1)$.

Exercise 3.6. Construct a sequence such that $X_n = O_P(1/n)$ and $\mathbb{E}[X_n] \neq O(1/n)$.

Exercise 3.7. Construct a sequence such that $X_n = O_P(1/n)$ and $P(|X_n| > u) \neq O(1/n)$ for every fixed $u > 0$.

3.2. Two Limit Theorems for Averages

There are three fundamental limit theorems for averages in probability theory: the law of large numbers (LLN), the central limit theorem (CLT), and the law of iterated logarithm (LIL). In a nutshell, the LLN states that the average converges to the mean; the CLT states that the convergence rate is $1/\sqrt{n}$ and the shape of the deviation approaches normality where the mean and variance are inherited from the elements of the average; the LIL states that the maximum deviation from the mean is eventually bounded by a vanishing function of n (Figure 3.2). The LLN and CLT are of great importance in statistics. The LIL, on the other hand, while proven useful in number theory, is not so much of interest to statisticians [**vdV98**, Section 2.7], hence omitted.

Hereafter we denote the sample average by $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$.

3.2.1. Laws of large numbers. There are two versions of the law of large numbers, the weak and the strong. The strong concludes almost sure convergence, and the weak concludes convergence in probability. As we discussed in Section 3.1, convergence in probability is enough for us, so the strong LLN is not of our concern.

Theorem 3.4 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random vectors. If $\mathbb{E}[\|X\|] < \infty$, then $\bar{X}_n \to^p \mathbb{E}[X]$.*

Proof. [**vdV98**, Proposition 2.16] and the remark thereunder. ■

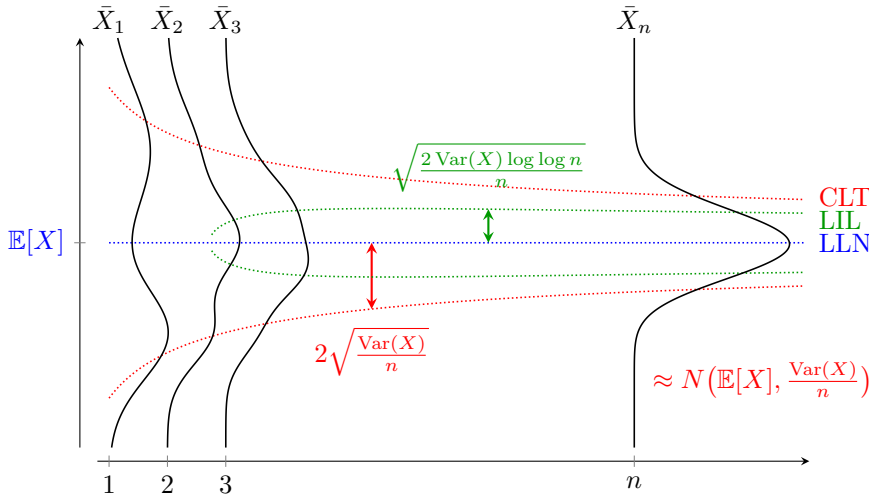


FIGURE 3.2. Three fundamental limit theorems. The LIL bound (green line) does not even exceed the 95% prediction interval by the CLT (red line) until much larger n.

Remark 3.5. Note that $\mathbb{E}[\|X\|] < \infty$ requires only the first absolute moment of each component of X even if $\|\cdot\|$ is a Euclidean norm. To see why, observe that for $X = (X_1, X_2)$, $\|X\| = \|(X_1, 0) + (0, X_2)\| \leq \|(X_1, 0)\| + \|(0, X_2)\| = |X_1| + |X_2|$.

The LLN upvotes the use of an average as the key summary statistic in various applications. For example, if an airline company wants to predict whether *one* passenger shows up to a booked flight, it can be highly uncertain; however, if they want to predict the percentage of no-shows among *all* passengers in one flight, it is much less uncertain and would be very close to the probability of a no-show, thanks to the LLN. Another example, a long-term investor may decide on their portfolio based on the average historical performance instead of the performance of the very last period, as the average would be a more precise predictor of the future performance in the long run.

3.2.2. Central limit theorem. The LLN does not tell us how close the average is to the expectation for a finite n. This is not always practical since we sometimes want to quantify (i) *how fast* the average converges to the expectation and (ii) *how* the average converges to the expectation. The CLT refines the LLN in both ways when a stronger condition can be imposed. Surprisingly, it turns out that a very mild assumption not only gives the rate of convergence but also characterizes the distribution of the average.

In particular, when each element X_i has a variance, the CLT first states that the distribution of an average $\bar{X}_n$ shrinks at the same speed as $1/\sqrt{n}$. This gives a rough idea of how close we can expect $\bar{X}_n$ is to $\mathbb{E}[X]$ for a finite n. However, there can be many ways in which something can converge to something at the speed of $1/\sqrt{n}$ (Figure 3.3). What is astonishing about the CLT is that, without any more additional assumptions, it states that $\bar{X}_n$ converges in the same way a sequence

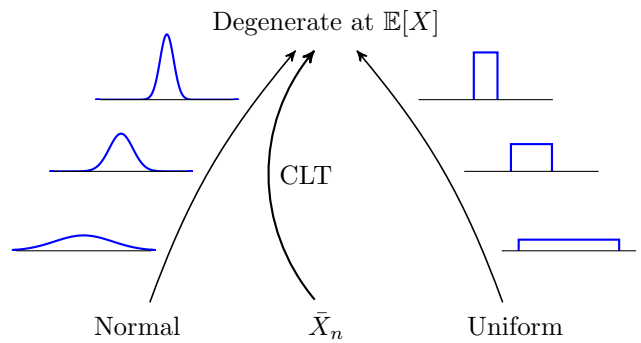


FIGURE 3.3. There are many ways for a distribution to shrink to degeneracy, but the CLT asserts that the distribution of an average consorts along with the normal sequence that goes degenerate at rate $1/\sqrt{n}$.

of normal distributions converges to $\mathbb{E}[X]$. In a more precise (but still informal) expression, we may understand this as

$$\bar{X}_n \rightsquigarrow N\Big(\mathbb{E}[X], \tfrac{1}{n}\operatorname{Var}(X)\Big).$$

Note that the variance of the normal distribution shrinks with $1/n$. Since the variance is in the squared units, $\bar{X}_n$ itself shrinks with $1/\sqrt{n}$.

Although this formulation neatly captures the intuition of the phenomenon, it is mathematically a bit bizarre in the sense that the "limit" on the RHS is a moving target that still depends on n. To make the variance independent of n, we may multiply both sides by $\sqrt{n}$, so that

$$\sqrt{n}\bar{X}_n \rightsquigarrow N(\sqrt{n}\mathbb{E}[X], \operatorname{Var}(X)).$$

Then again, the mean now depends on n, so let's subtract $\sqrt{n}\mathbb{E}[X]$ from both sides, and we obtain the "mathematically correct" version of the statement.

Theorem 3.5 (Central limit theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random vectors with $\mathbb{E}[XX'] < \infty$. Then*

$$\sqrt{n}(\bar{X}_n - \mathbb{E}[X]) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}[X]) \rightsquigarrow N(0, \operatorname{Var}(X)).$$

PROOF. [**vdV98**, Proposition 2.17 and Example 2.18]. ■

The central limit theorem is arguably the most important theorem in asymptotic statistics. Historically, the name was given to indicate that it played the central role in the collection of limit theorems ("central" limit theorem), although some French authors now call it *le théorème de la limite centrale* ("central limit" theorem).

Note that the only essential assumption of the CLT is the existence of the second moment; we make no assumption on the *shape* of the distribution of each X_i, yet we have as a conclusion the *shape* of $\bar{X}_n$. Thus, the distribution of an average is a consequence, not an assumption. It frees us from making a strong distributional assumption on the data and enables us to do valid inference for which the knowledge

of distribution is imperative. The cost, on the other hand, is the largeness of the sample, which is often tolerable in economic applications.

Example 3.4 (Insurance). Let us say that the car accident happens with probability 1%. If an accident happens, the driver incurs a loss of 100 (in whatever units); if no accident occurs, she incurs no loss. In particular, the distribution of the loss X for a driver is $P(X = 0) = 99\%$ and $P(X = 100) = 1\%$. If the driver wants to hedge the risk by maintaining the liquid asset worth the maximum 99.9% loss (i.e., by controlling the VaR at 99.9%), then she must keep the amount of 100 as a dormant, unusable money in her bank account.

If an insurance company covers the loss of n independent drivers, then the distribution of the loss (total insurance claim) $n\bar{X}_n$ is approximately $N(n\mathbb{E}[X], n\operatorname{Var}(X)) = N(n, 99n)$ by the CLT (and Slutsky's lemma). This company's 99.9% VaR is $n + 3.09\sqrt{99n} \approx n + 30.75\sqrt{n}$, so they can hedge the risk by collecting the premium of at least $1 + 30.75/\sqrt{n}$. For $n = 100$, this value is 1.31 and, for $n = 1{,}000$, it goes as low as 1.03. This means that the more insureds the insurance company has, the lower the premium it can set and the more efficient the risk sharing becomes.

From a driver's perspective, paying the insurance premium of just over 1 frees up her asset worth 100 and covers her entire risk associated with car accidents (at least in pecuniary terms), so insurance sounds like a great economic institution. From an economist's perspective, however, the story looks different; since larger insurance companies can offer lower premiums, this seems to suggest that the market is always driven toward monopoly, which comes with all the bad consequences. Can we avoid this? How do we sustain competition while encouraging risk sharing?

Exercise 3.8. Instead of an arithmetic average, consider $\tilde{X}_n = (X_1 X_2 \cdots X_n)^{1/\sqrt{n}}$, where $X_1, \ldots, X_n$ are i.i.d. positive random variables with $\mathbb{E}[(\log X_i)^2] < \infty$. Derive the asymptotic distribution of $\tilde{X}_n$. *Hint: Take the logarithm and apply the CLT.*

Remark 3.6. While convergence in distribution itself does not imply convergence of moments (Exercise 3.4), the condition of the CLT does imply that $\operatorname{Var}(\sqrt{n}\bar{X}_n)$ converges to $\operatorname{Var}(X)$; in fact, $\operatorname{Var}(\sqrt{n}\bar{X}_n) = \operatorname{Var}(X)$ for every n.

Remark 3.7. There are many variants of the CLT with relaxed assumptions.

- The CLT for independent but not identically distributed (i.n.i.d.) random variables with finite variances is known as the Lindeberg–Feller CLT [**vdV98**, Proposition 2.27] and is used in cross-sectional econometrics.
- There are CLTs for time seires dependence with finite variance [**Dav21**, Chapter 25], which are used in time series econometrics.
- There is also a CLT that does not assume a particular dependence structure [**Dav21**, Theorem 25.1].
- When the variance is infinite, the generalized CLT asserts that the average converges to an α-stable distribution [**Dav21**, Theorem 24.23]. The α-stable distributions occasionally appear in the research of fat tails in finance.
- When an average of random functions is of interest, there may be a functional CLT that concludes convergence to a Gaussian process. This convergence is

often considered under a uniform metric, in which case it is specifically called the uniform CLT [Dud14].

Exercise 3.9 (CLT and dependence). Although the assumption of i.i.d. observations can be relaxed, zero pairwise correlation is certainly not enough for a CLT. Construct a sequence $X_1, X_2, \dots$ such that $\mathbb{E}[X_i] = 0$, $\mathrm{Var}(X_i) = \sigma^2$, and $\mathrm{Cov}(X_i, X_j) = 0$ for every $i \neq j$, but $\sqrt{n}\bar{X}_n$ does not converge in distribution to a normal.

The CLT is still an approximation, and one can go one more step and ask "but how close is the distribution of the average to the normal distribution?" The Berry–Esseen theorem can bound the deviation from the normal distribution, which is sometimes used in finance.

Proposition 3.6 (Berry–Esseen)**.** *Denote by Φ the cdf of a standard normal distribution. Let $X_1, X_2, \dots$ be i.i.d. univariate random variables with $\mathbb{E}[|X|^3] < \infty$ and $\bar{Z}_n := \sqrt{n}\,\mathrm{Var}(X)^{-1/2}(\bar{X}_n - \mathbb{E}[X])$. Then, there exists a constant $0.4 < C < 0.5$ such that*

$$\sup_{x\in\mathbb{R}} |F_{\bar{Z}_n}(x) - \Phi(x)| \leq \frac{C}{\sqrt{n}}\,\mathbb{E}\left[\left|\frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}\,(X)}}\right|^3\right].$$

Proof. A stronger version is proved in [SS93, Theorem 3.5.2]. ■

3.3. The Delta Method

Sometimes a quantity of interest is not an average per se but a transformation thereof. Fortunately, convergence towards normality is preserved under smooth transformation. In particular, a smooth transformation of an average, $g(\bar{X}_n)$, converges in distribution to a normal distribution. This is because the average $\bar{X}_n$ converges to a tighter and tighter normal distribution by the CLT, and a smooth transformation of a shrinking quantity approaches a linear transformation, so the normality is preserved under a linear operation (Figure 3.4A). Conceptually, this is a continuous mapping theorem (Theorem 3.3) where the map becomes closer and closer to a linear map.

Theorem 3.7 (Delta method)**.** *Let X_i be a k-dimensional random vector and $g : \mathbb{R}^k \to \mathbb{R}^\ell$ be differentiable at $\mathbb{E}[X]$ with an $\ell \times k$ Jacobian $g'(\mathbb{E}[X])$. If $\bar{X}_n$ follows the CLT, then*

$$\sqrt{n}(g(\bar{X}_n) - g(\mathbb{E}[X])) \rightsquigarrow N\Big(0, g'(\mathbb{E}[X])\,\mathrm{Var}(X)g'(\mathbb{E}[X])'\Big).$$

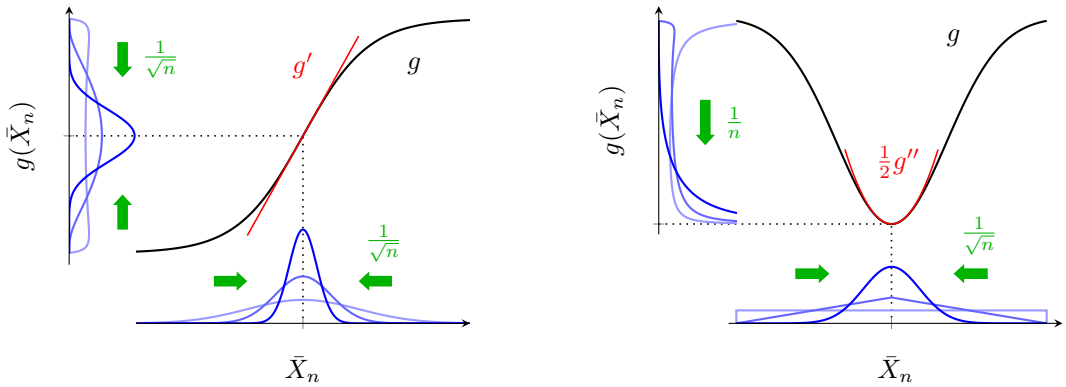
Proof. By Taylor's theorem,

$$\sqrt{n}(g(\bar{X}_n) - g(\mathbb{E}[X])) = g'(\mathbb{E}[X])\sqrt{n}(\bar{X}_n - \mathbb{E}[X]) + \sqrt{n}\,o(\|\bar{X}_n - \mathbb{E}[X]\|),$$

where $o(\cdot)$ satisfies $\frac{1}{t}o(t) \to 0$ as $t \to 0$. So, the remainder vanishes as

$$\sqrt{n}\|\bar{X}_n - \mathbb{E}[X]\| \cdot \frac{o(\|\bar{X}_n - \mathbb{E}[X]\|)}{\|\bar{X}_n - \mathbb{E}[X]\|} = O_P(1) \cdot o_P(1) = o_P(1).$$

Slutsky's lemma (Theorem 3.2 (iv)) implies that the first term converges in distribution to $N(0, g'(\mathbb{E}[X])\,\mathrm{Var}(X)g'(\mathbb{E}[X])')$. ■



(A) If $g'(\mathbb{E}[X]) \neq 0$, $g(\bar{X}_n)$ approaches normal around $g(\mathbb{E}[X])$ at rate $1/\sqrt{n}$ as $\bar{X}_n$ approaches normal around $\mathbb{E}[X]$ at rate $1/\sqrt{n}$.

(B) If $g'(\mathbb{E}[X]) = 0$ and $g''(\mathbb{E}[X]) \neq 0$, $g(\bar{X}_n)$ approaches χ^2 around $g(\mathbb{E}[X])$ at rate $1/n$ as $\bar{X}_n$ approaches normal around $\mathbb{E}[X]$ at rate $1/\sqrt{n}$.

FIGURE 3.4. The delta method.

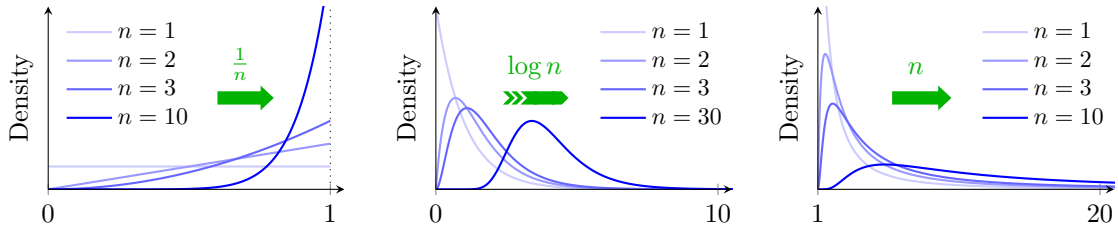
Remark 3.8. Theorem 3.7 holds even when $g'(\mathbb{E}[X]) = 0$ in the sense that $\sqrt{n}(g(\bar{X}_n) - g(\mathbb{E}[X]))$ converges in distribution to 0. However, to extract a nondegenerate distribution, we need to look at a higher-order term. For example, $n(g(\bar{X}_n) - g(\mathbb{E}[X]))$ may converge in distribution to (a multiple of) a chi-square distribution if the second derivative exists and is nonzero (Figure 3.4B).

EXERCISE 3.10. Related to Exercise 3.8, but now consider the geometric average $\breve{X}_n = (X_1 X_2 \cdots X_n)^{1/n}$. Find the asymptotic distribution of $\breve{X}_n$.

EXERCISE 3.11. Let $X_1, X_2, \ldots$ be i.i.d. and each have the pdf $p(x) = \frac{1}{2}x^2 e^{-x} \mathbb{1}\{x > 0\}$. Under appropriate scaling, derive the asymptotic distributions of the arithmetic average $(X_1 + \cdots + X_n)/n$, the geometric average $(X_1 \cdots X_n)^{1/n}$, the harmonic average $n(X_1^{-1} + \cdots + X_n^{-1})^{-1}$, and the quadratic average $\sqrt{(X_1^2 + \cdots + X_n^2)/n}$.

3.4. Extreme Value Theory for Extremes

The limit theorems for averages are extremely powerful as many quantities we encounter in statistics are represented as averages. However, some applications do involve quantities other than averages. A notable example is the *extremes*. For instance, the financial risk is sometimes measured by the Value-at-Risk, which is defined as the maximal loss the investor may incur with some large probability, say 99%. In the sample of historical returns, this corresponds to the 99% percentile of the loss distribution. In economics, some auction data may only contain the winning bids, i.e., the highest bids of all bids submitted to respective auctions. In meteorology, the temperature may be recorded as the daily maximum and minimum. Such "extreme" values cannot be represented as averages, and as such their behaviors may differ significantly from averages.



(A) Maximum of n standard uniform random variables shrinks to the boundary at rate $1/n$.

(B) Maximum of n standard exponential random variables glides off to infinity at rate $\log n$.

(C) Maximum of n unit Pareto random variables expands to infinity at rate n.

FIGURE 3.5. Three types of convergence for extremes.

There is a branch of probability theory called the *extreme value theory (EVT)*, which deals with the approximation of such quantities. As surprising as it is, even the extreme observation (whose value is essentially determined by *one* observation) elicits regularity when the dataset out of which it is picked is large. In this sense, the EVT is the approximation theory for "rare" events when the CLT is the approximation theory for "ordinary" events. Just like an average (mostly) distributes as a normal distribution, an extreme distributes as a *generalized extreme value (GEV) distribution*, which consists of three types of distributions.[1]

The simplest example to illustrate how it happens is the maximum of i.i.d. observations. We use the following notation. For univariate i.i.d. random variables $X_1, \ldots, X_n$, we denote their rearrangement into an ascending order by $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. We call $X_{(1)}, \ldots, X_{(n)}$ the *order statistics*. Therefore, $X_{(n)}$ is the maximum of n observations. Now, the maximum of bounded random variables tends to converge to a type III distribution, as the following example illustrates.

EXAMPLE 3.5 (Maximum of uniform variables). The cdf of the maximum $X_{(n)}$ of n independent standard uniform variables is given by $F_{X_{(n)}}(x) = P(X_1 \leq x, \ldots, X_n \leq x) = P(X_1 \leq x) \cdots P(X_n \leq x) = x^n$ for $0 \leq x \leq 1$. Now, consider the linear transformation $Y = (X_{(n)} - b)/a$ for some $a > 0$ and b. The cdf of Y is $F_Y(y) = F_X(b + ay) = (b + ay)^n = b^n(1 + ay/b)^n \to b^n \exp(nay/b)$. Therefore, if we set $a = 1/n$ and $b = 1$, the cdf of Y converges pointwise to $\exp(y)$ for $y \leq 0$. In light of Theorem 3.1, Y converges in distribution to the reversed standard exponential distribution, which is a special case of the Weibull (extreme value type III) distribution. Note that this also implies that $X_{(n)} - 1 = O_P(1/n)$, so $X_{(n)}$ converges much faster than an average would (Figure 3.5A).

The maximum of exponentially tailed random variables tends to converge to a type I distribution. The type I distribution is also used as a foundation for logistic regression (Section 8.3.2).

[1]Interestingly, an intermediate quantile such as a median is also determined by the value of one observation but asymptotes to a normal distribution (Example 4.1).

EXAMPLE 3.6 (Maximum of exponential variables). The cdf of the maximum of n independent standard exponential variables is $F_{X_{(n)}}(x) = (1-e^{-x})^n$ for $x \geq 0$. The cdf of $Y = (X_{(n)} - b)/a$ for $a > 0$ is $F_Y(y) = [1-\exp(-b-ay)]^n \to \exp(-n\exp(-b-ay))$. If we set $a = 1$ and $b = \log n$, then F_Y converges pointwise to $\exp(-\exp(-y))$. Thus, Y converges in distribution to the standard Gumbel distribution, which is an extreme value type I distribution. In this case, $X_{(n)}$ glides off to infinity at rate $\log n$, but $X_{(n)} - \log n$ stays $O_P(1)$ (Figure 3.5B).

The maximum of fat-tailed (polynomially tailed) random variables tends to converge to a type II distribution.

EXAMPLE 3.7 (Maximum of Pareto variables). The cdf of the maximum of n independent unit Pareto variables is $F_{X_{(n)}}(x) = (1 - 1/x)^n$ for $x \geq 1$. The cdf of $Y = (X_{(n)} - b)/a$ for $a > 0$ is $F_Y(y) = [1 - 1/(b + ay)]^n \to \exp(-n/(b + ay))$. If we set $a = n$ and $b = 0$, then F_Y converges pointwise to $\exp(-1/y)$ for $y > 0$. Thus, Y converges in distribution to the unit Fréchet distribution, which is an extreme value type II distribution. Here, $X_{(n)}$ bloats at rate n, so $X_{(n)}/n = O_P(1)$ (Figure 3.5C).

EXERCISE 3.12. Derive the asymptotic distribution for the maximum of independent Pareto variables with parameter 2 that have the cdf $F_X(x) = 1 - 1/x^2$ for $x \geq 1$.

Just as the CLT does, the EVT allows us to conduct inference on extremes without making much distributional assumptions on individual observations. The detailed treatment of the EVT is out of the scope of this course. I refer the interested reader to [**EKM97**].

CHAPTER 4

Principles of Estimation

Measure what is measurable, and make measurable what is not so.

Galileo Galilei, quoted in galilée by thomas henri martin, 1868

Statistical analysis is depicted in Figure 4.1. It starts with *modeling*, which specifies the parameter θ and the model $\mathcal{P}$. The *model* is the set $\mathcal{P}$ of candidate probability distributions we deem possible, and the *parameter* is some characteristic θ of the probability distribution. In this sense, θ can be regarded as a function defined on $\mathcal{P}$. The set of all values of θ spanned by $\mathcal{P}$ is denoted by Θ and is called the parameter space. The parameter may be split into two parts: the parameter of interest and the *nuisance parameter*, which we use for the sake of modeling but are not interested in knowing. After the modeling stage, we observe the sample X from an unknown probability distribution and infer θ to conclude the analysis. This last step is divided into two parts: estimation and inference.

Estimation is the process of formulating a guess $\hat{\theta}$ on θ as a function of X. The map $X \mapsto \hat{\theta}$ is called the *estimator*, and the specific value of $\hat{\theta}$ calculated with the realized value of X is called the *estimate*. The goal of estimation includes constructing an estimator that works under reasonable assumptions and having as precise an estimator as possible. *Inference* is the process of assessing how much information X contains about θ, such as quantifying the uncertainty of $\hat{\theta}$ and making a judgement about a hypothesis on θ. The goal of inference includes yielding an interpretable result on which to take action and drawing as strong a conclusion as possible. Usually,

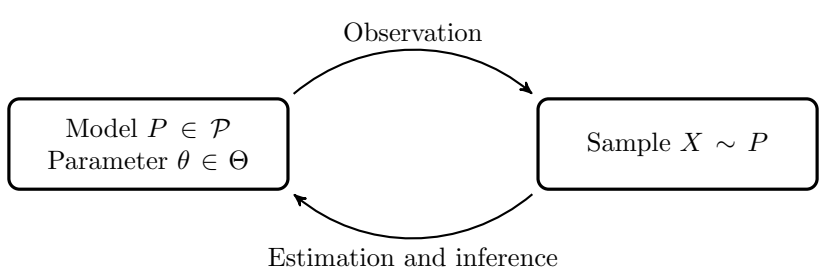


Figure 4.1. Statistical analysis.

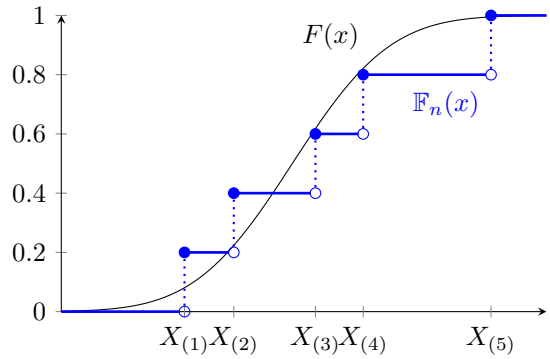


FIGURE 4.2. Empirical and population cumulative distribution functions.

estimation precedes inference, but some applications consist solely of estimation or solely of inference.

4.1. Construction of Estimators

As stated earlier, the parameter of interest is given as a map $\theta : \mathcal{P} \to \Theta$ from the model $\mathcal{P}$ to the parameter space Θ. In most cases, Θ is (a subset of) a Euclidean space, but it can also be a space of functions or more complicated objects.

A natural point to start the discussion is the case where θ is an identity, that is, $\Theta = \mathcal{P}$, so the probability distribution itself is our target.

4.1.1. Empirical distribution. When we observe i.i.d. univariate random variables $X_1, \ldots, X_n$ from an unknown distribution P, a naive yet powerful guess about P is that it takes values $\{X_1, \ldots, X_n\}$ with equal probability. This guess is called the *empirical distribution* (or *empirical measure*) and is denoted by $\mathbb{P}_n$. The cdf of $\mathbb{P}_n$ is called the *empirical distribution function* and takes the form

$$\mathbb{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}.$$

Note that this is not continuous (let alone absolutely continuous), so the empirical distribution does not have a pdf (Figure 4.2).

The empirical distribution function has the nice property that at every point x, its value $\mathbb{F}_n(x)$ converges to a normal distribution centered at the true value $F(x)$ with variance $F(x)[1 - F(x)]$. This is not even hard to prove.

Theorem 4.1 (Convergence of empirical distribution)**.** *For every F and every $x \in \mathbb{R}$,*

$$\sqrt{n}(\mathbb{F}_n(x) - F(x)) \rightsquigarrow N(0, F(x)[1 - F(x)]).$$

PROOF. Observe that $\mathbb{1}\{X_i \le x\}$ is a Bernoulli random variable that takes 1 with probability $F(x)$. So, $\mathbb{E}[\mathbb{1}\{X_i \le x\}] = F(x)$ and $\mathrm{Var}(\mathbb{1}\{X_i \le x\}) = F(x)[1 - F(x)]$. Then, the claim follows from the CLT. ■

(A) Pointwise 50% confidence band constructed with Theorem 4.1.

(B) Uniform 50% confidence band constructed with Proposition 4.3.

FIGURE 4.3. Asymptotic pointwise and uniform confidence bands for F.

Indeed, much stronger results are known to hold. Remind us that the model $\mathcal{P}$ here is the set of all univariate distribution functions and the parameter space is the set of all univariate cdfs, say $\mathcal{F}$. Therefore, we are estimating a *function*. Now, if we measure the distance of two functions by the maximum vertical distance, that is, by

$$\|f - g\|_\infty := \sup_{x \in \mathbb{R}} |f(x) - g(x)|,$$

it is known that convergence as in Theorem 4.1 still holds.

Proposition 4.2 (Glivenko–Cantelli; [**vdV98**, Theorem 19.1])**.** *For every F,*

$$\|\mathbb{F}_n - F\|_\infty \xrightarrow{\text{as}} 0.$$

Proposition 4.3 (Donsker; [**vdV98**, Theorem 19.3])**.** *For every F,*

$$\sqrt{n}(\mathbb{F}_n - F) \rightsquigarrow \mathbb{G}_F \quad \textit{in} \quad \|\cdot\|_\infty,$$

where $\mathbb{G}_F$ is a Gaussian process with a mean function identically 0 *and a covariance function* $\mathrm{Cov}(\mathbb{G}_F(x), \mathbb{G}_F(y)) = F(x \wedge y)[1 - F(x \vee y)]$.

The random function $\sqrt{n}(\mathbb{F}_n - F)$ is called the *empirical process*. These results can be used to construct a uniform confidence band for F (Figure 4.3) or to construct a test of a hypothesis about the entire functional form of F (*Kolmogorov–Smirnov test*).

What's nice about the empirical distribution is that it captures all of the informative randomness of i.i.d. observations (in technical terms, it is a *sufficient statistic* for P). So, almost any statistic (i.e., almost any function of the data) can be written as a function of the empirical distribution, to which the delta method might be applicable.

Proposition 4.4 (Functional delta method; [**vdV98**, Theorem 20.8])**.** *Let $\phi : \mathcal{F} \to \mathbb{R}$ be a functional defined on a set of cdfs. If ϕ is differentiable in a suitable sense with the derivative map ϕ'_F, then*

$$\sqrt{n}(\phi(\mathbb{F}_n) - \phi(F)) \rightsquigarrow \phi'_F(\mathbb{G}_F).$$

In many cases, this is an overkill way to derive the distribution of a finite-dimensional estimator, but sometimes this is the only (or most general) way. These results can be extended to multivariate cases and more.

4.1.2. Plug-in estimators. When θ is not an identity, a straightforward way to construct an estimator is to plug in the empirical distribution. Recall that θ is a function defined on the model $\mathcal{P}$. If θ is also defined at $\mathbb{P}_n$ (either because $\mathcal{P}$ includes $\mathbb{P}_n$ or θ can be extended to $\mathbb{P}_n$), then $\hat{\theta} := \theta(\mathbb{P}_n)$ is a natural estimator for θ.

Example 4.1 (Descriptive statistics). The mean $\mathbb{E}[X]$ can be regarded as a parameter $P \mapsto \int x dP(x)$. Thus, the plug-in estimator of the mean is the *sample average* $\mathbb{E}_n[X] = \int x d\mathbb{P}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$. If X_i has variance, the CLT implies that $\sqrt{n}(\bar{X}_n - \mathbb{E}[X]) \rightsquigarrow N(0, \mathrm{Var}(X))$.

Similarly, the variance $\mathrm{Var}(X)$ and covariance $\mathrm{Cov}(X, Y)$ are given as parameters $P \mapsto \int (x - \int x dP)^2 dP$ and $P \mapsto \int (x - \int x dP)(y - \int y dP) dP$. So, their plug-in estimators are $\int (x - \int x d\mathbb{P}_n)^2 d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $\int (x - \int x d\mathbb{P}_n)(y - \int y d\mathbb{P}_n) d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$. If X and Y have the fourth moments, it is straightforward to compute their asymptotic distributions using the CLT.

The αth quantile can be seen as a parameter $P \mapsto F^{-1}(\alpha)$, where F^{-1} denotes the left-continuous generalized inverse of the cdf as in Definition 2.2. Its plug-in estimator is the *sample αth quantile* $\mathbb{Q}_n(\alpha) = \mathbb{F}_n^{-1}(\alpha) = \inf\{x \in \mathbb{R} : \mathbb{F}_n(x) \geq \alpha\}$. If F has a positive density at $F^{-1}(\alpha)$, then the asymptotic distribution of $\sqrt{n}[\mathbb{F}_n^{-1}(\alpha) - F_X^{-1}(\alpha)]$ can be calculated using Proposition 4.4 as $N(0, \alpha(1-\alpha)/p_X \circ F_X^{-1}(\alpha)^2)$ [**vdV98**, Example 20.5].

Exercise 4.1 (Quantile function and order statistics). Let $X_1, \ldots, X_n$ be univariate i.i.d. random variables. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the rearrangement of $X_1, \ldots, X_n$ in an ascending order, so $X_{(k)}$ is the kth smallest observation. We call $X_{(1)}, \ldots, X_{(n)}$ the *order statistics*. Show that the empirical quantile function is given by $\mathbb{Q}_n(u) = X_{(\lceil nu \rceil)}$, where $X_{(k)}$ is the kth smallest observation as introduced in Section 3.4 and $\lceil a \rceil$ is the smallest integer greater than or equal to a.

Example 4.2 (Z-estimation). A parameter may be given as the zero of an equation $\psi(P, \theta) = 0$. For example, the parameters of a consumer's utility function may be given as the values that satisfy the Euler equation. In this case, the plug-in estimator is the solution to $\psi(\mathbb{P}_n, \theta) = 0$. The condition $\psi(P, \theta) = 0$ is called the *moment condition* in econometrics. For a specific example, consider the capital asset pricing model (CAPM). It decomposes the excess return of a stock Y into the market excess return X and the return orthogonal to it $\alpha + \varepsilon$, that is, $Y = \alpha + \beta X + \varepsilon$ for some (α, β) such that $\mathbb{E}[X\varepsilon] = 0$ and $\mathbb{E}[\varepsilon] = 0$. In other words, the parameter $\theta = (\alpha, \beta)$ is defined by the zero of $\mathbb{E}[X(Y - \alpha - \beta X)] = 0$. The plug-in estimator finds the value of $(\hat{\alpha}, \hat{\beta})$ such that $\mathbb{E}_n[X_i(Y_i - \hat{\alpha} - \hat{\beta} X_i)] = \frac{1}{n}\sum_{i=1}^n X_i(Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$ and $\mathbb{E}_n[Y_i - \hat{\alpha} - \hat{\beta} X_i] = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$, which yields $\hat{\beta} = \widehat{\mathrm{Cov}}(X, Y)/\widehat{\mathrm{Var}}(X)$ and $\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{X}_n$. The asymptotic distribution of $(\hat{\alpha}, \hat{\beta})$ follows from Theorem 7.1. More generally, there are generic theorems to derive the asymptotic distribution of a Z-estimator [**vdV98**, Chapter 5].

EXAMPLE 4.3 (M-estimation). Another prevalent situation is that a parameter of interest is given as the maximizer, $\theta(P) = \arg\max_\vartheta M(P, \vartheta)$. Then, the plug-in estimator is given by the maximizer of the sample objective function $\hat{\theta} = \arg\max_\vartheta M(\mathbb{P}_n, \vartheta)$. If M is differentiable in a suitable sense in θ, this reduces to a Z-estimation by taking the first-order condition (FOC), $\frac{\partial}{\partial\theta} M(\mathbb{P}_n, \hat{\theta}) = 0$.[1] M-estimation is also called *extremum estimation*. A concrete example is the conditional Value-at-Risk (VaR) estimation. Let Y be the return of the portfolio and X be the information available for prediction, such as the lagged returns or the values of other assets. We are interested in the τth quantile of Y conditional on X, which is a solution to $\min_f \mathbb{E}[\rho_\tau(Y - f(X))]$ where ρ_τ is the check function for the τth quantile (recall Theorem 2.6) [**Koe05, KCHP18**]. Then, the plug-in estimator of the conditional VaR is a solution to $\min_f \mathbb{E}_n[\rho_\alpha(Y_i - f(X_i))]$ over some class of functions. There are general theorems to derive the asymptotic distribution of an M-estimator [**vdV98**, Chapter 5].

The word "plug-in estimator" is by no means reserved for plugging in the empirical distribution. For example, if a parameter is identified by $\psi(\theta, \eta) = 0$ and η is estimated by some method, the solution θ for $\psi(\theta, \hat{\eta}) = 0$ in is also called a plug-in estimator.

4.1.3. Other estimators. There are various reasons why we may use estimators other than the plug-in. Sometimes the plug-in estimator does not exist; at other times, a different type of an estimator has a nicer property than the plug-in.

EXAMPLE 4.4 (Bessel's correction). Denote $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}(X)$. Some may argue that the plug-in estimator for σ^2 is not desirable as it underestimates σ^2,

$$\begin{aligned}
\mathbb{E}[\tfrac{1}{n}\textstyle\sum_i (X_i - \bar{X})^2] &= \tfrac{1}{n}\textstyle\sum_i \mathbb{E}[(X_i - \mu - \bar{X} + \mu)^2] \\
&= \tfrac{1}{n}\textstyle\sum_i \mathbb{E}[(X_i - \mu)^2] - 2\tfrac{1}{n}\textstyle\sum_i \mathbb{E}[(X_i - \mu)(\bar{X} - \mu)] + \mathbb{E}[(\bar{X} - \mu)(\bar{X} - \mu)] \\
&= \sigma^2 - \tfrac{2}{n^2}\textstyle\sum_i\sum_j \mathbb{E}[(X_i - \mu)(X_j - \mu)] + \tfrac{1}{n^2}\textstyle\sum_j\sum_k \mathbb{E}[(X_j - \mu)(X_k - \mu)] \\
&= \sigma^2 - \tfrac{2}{n}\sigma^2 + \tfrac{1}{n}\sigma^2 = \tfrac{n-1}{n}\sigma^2.
\end{aligned}$$

So, the plug-in estimator on average gives an estimate smaller by the factor of $\frac{n-1}{n}$. On the other hand, the adjusted estimator $\widehat{\mathrm{Var}}(X) = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$ has expectation equal to σ^2. This is usually what is called the *sample variance*. For the same reason, we use the *sample covariance* $\widehat{\mathrm{Cov}}(X, Y) = \frac{1}{n-1}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})$ as our default covariance estimator. Since the difference of n-scaling and $(n-1)$-scaling vanishes asymptotically, it follows by Slutsky's lemma that the asymptotic distributions of these are the same as the plug-in estimators in Example 4.1.

EXERCISE 4.2. Show that the square root of the sample variance $\sqrt{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2}$ underestimates the standard deviation. *Hint: Use Jensen's inequality.*

[1] Conversely, a Z-estimation problem $\psi(P, \theta) = 0$ can be made into an M-estimation by writing $\theta = \arg\min_\vartheta \|\psi(P, \theta)\|^2$.

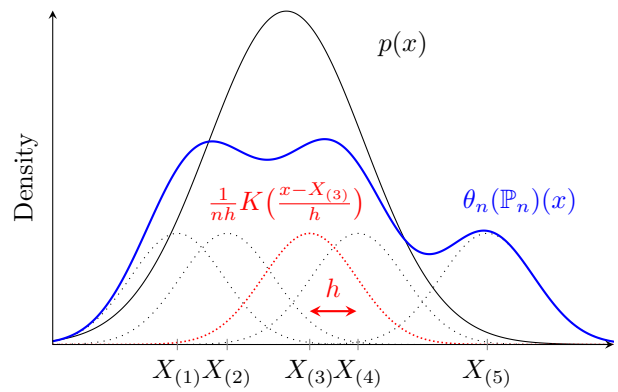


FIGURE 4.4. Kernel density estimation. If the bandwidth h is sent to 0 at an appropriate rate relative to n, the kernel density estimator converges to the true density.

EXAMPLE 4.5 (Kernel density estimation). A canonical example in which a simple plug-in estimator does not exist is the density estimation, which appear, e.g., in auction research [**GPV00**]. Let $\theta(P)$ be the pdf of P. Then $\theta(\mathbb{P}_n)$ does not exist as the empirical distribution is a step function. One way to estimate the pdf is the kernel density estimator

$$\theta_n(\mathbb{P}_n)(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big),$$

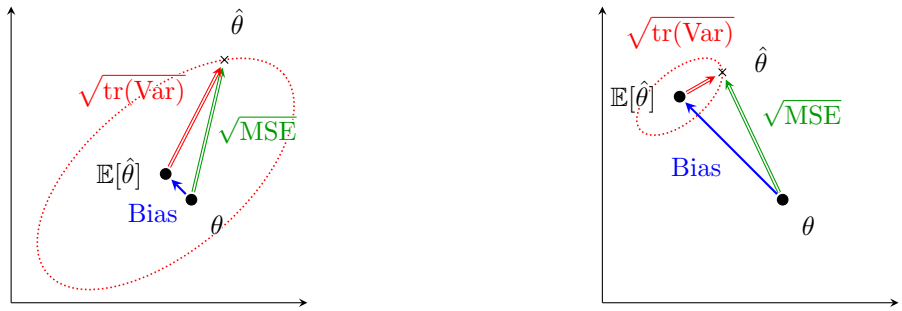
where K is some nice function that integrates to one (often is itself a pdf), and h is a tuning parameter that converges slow enough to 0 as $n \to \infty$ so that nh still diverges [**vdV98**, Chapter 24]. Figure 4.4 illustrates this method.

EXAMPLE 4.6 (Bayes estimator). A Bayesian decision maker has a prior $\pi(\theta)$ on θ and the model $X \sim P_\theta$. The model can be interpreted as the conditional distribution of the data X given θ, so the product $\pi(\theta)p_\theta(x)$ gives the joint pdf of the parameter and the data (θ, X). When she observes X, she updates her belief about θ using the conditional distribution of θ given X, called the posterior,

$$\pi(\theta \mid X) = \frac{\pi(\theta)p_\theta(x)}{\int \pi(\theta)p_\theta(x)d\theta}.$$

She may then use this posterior to minimize her loss or risk [**LC98**, Chapter 4] or calculate the posterior mode or mean to obtain a point estimate. Bayesian inference is foundational in decision theory. Statisticians also find Bayes estimators appealing as appropriately chosen priors lead to nice characteristics even from a frequentist's perspective [**vdV98**, Chapter 10].

EXAMPLE 4.7 (Bootstrap estimation). Some estimators are so complicated that we cannot analytically derive their distributions. However, if we know the true data generating process P, we can simulate a lot of samples $X \sim P$ on a computer and calculate $\hat{\theta}(X)$ many times to simulate the sampling distribution of $\hat{\theta}$. The *bootstrap*



(A) Accurate but not precise estimator. (B) Precise but not accurate estimator.

FIGURE 4.5. Decomposition $(\hat{\theta} - \theta) = (\mathbb{E}[\hat{\theta}] - \theta) + (\hat{\theta} - \mathbb{E}[\hat{\theta}])$. The single arrow indicates a nonrandom quantity; the double arrows indicate random quantities.

procedure replaces the unknown P with the empirical distribution $\mathbb{P}_n$. Thus, we draw a set of n observations X^* from $\mathbb{P}_n$ (so some observations may be drawn more than once while some others may not appear at all) and calculate $\hat{\theta}(X^*)$. We then repeat this process many times to "estimate" the sampling distribution of $\hat{\theta}$. This way we can obtain the sampling distribution purely computationally without having to take on the burden of doing any theory. For example, the standard error of $\hat{\theta}(X)$ can be estimated by the sample standard error of this collection of $\hat{\theta}(X^*)$. This obviously sounds too good to be true—using the data and only data to estimate their own inherent uncertainty. The name *bootstrap* comes from the fact that it smacks of pulling oneself up by one's own bootstraps.[2] Bootstrap is an example of more general *resampling* methods.

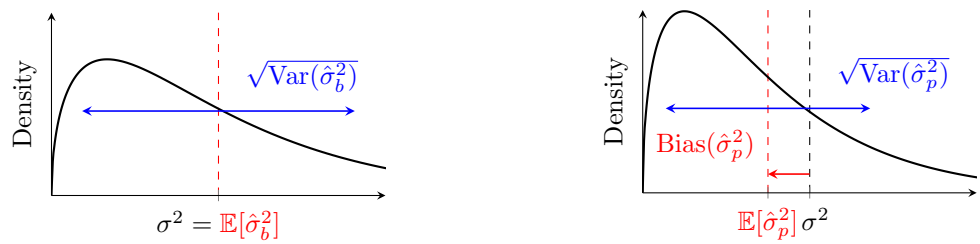
4.2. Bias and Variance of Estimators

In order to discuss which estimator to employ or how to construct a new estimator, we need to have clear desiderata. A natural criterion is to prefer an estimator that is close to the true parameter in terms of the mean squared error.

Definition 4.1 (Mean squared error)**.** The *mean squared error (MSE)* of $\hat{\theta}$ is given by $\text{MSE}(\hat{\theta}) \coloneqq \mathbb{E}[(\hat{\theta} - \theta)'(\hat{\theta} - \theta)]$.

There are two distinct channels through which the MSE can become large. The first is an inherent variation of $\hat{\theta}$; Figure 4.5A presents an estimator whose average location is close to θ ("accurate") but its random variation is huge. The second is the overall misposition of $\hat{\theta}$; Figure 4.5B shows an estimator whose variation is small ("precise") but its center is far off from θ. This motivates us to decompose the MSE into the measures of accuracy and precision, respectively called the bias and variance.

[2]In the original tale of *The Surprising Adventures of Baron Munchausen*, Munchausen pulls his own *pigtail* to drag himself out of a swamp, but "bootstrapping" is undeniably a better name than "pigtailing."



(A) Unbiased estimator $\hat{\sigma}_b^2$ has larger variance and larger MSE.

(B) Plug-in estimator $\hat{\sigma}_p^2$ has bias but in turn smaller variance and smaller MSE.

FIGURE 4.6. Distributions of $\hat{\sigma}_b^2$ and $\hat{\sigma}_p^2$ when $X_i \sim N(\mu, \sigma^2)$ and the bias-variance tradeoff.

Definition 4.2 (Bias and variance)**.** The *bias* of an estimator $\hat{\theta}$ of θ is defined by $\mathrm{Bias}(\hat{\theta}) := \mathbb{E}[\hat{\theta}] - \theta$. The *variance* of $\hat{\theta}$ is $\mathrm{Var}(\hat{\theta}) := \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\hat{\theta} - \mathbb{E}[\hat{\theta}])']$.

The variation of an estimator around the true value is then decomposed into the bias and the variance, thanks to the orthogonality of the second moment.

Theorem 4.5 (Decomposition of MSE)**.** $\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}(\hat{\theta})' \mathrm{Bias}(\hat{\theta}) + \mathrm{tr}(\mathrm{Var}(\hat{\theta}))$.

EXERCISE 4.3. Prove Theorem 4.5. *Hint: Use the add-and-subtract strategy.*

When we try to find an estimator that achieves a smaller MSE, we often encounter a situation in which improving the bias increases the variance and lowering the variance worsens the bias. This phenomenon is known as the *bias-variance tradeoff*. An illustrative example is the estimation of variance.

EXAMPLE 4.8 (Bias-variance tradeoff). Let $\hat{\sigma}_p^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$ and $\hat{\sigma}_b^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$. As seen in Example 4.4, $\hat{\sigma}_p^2$ underestimates σ^2 in that $\mathrm{Bias}(\hat{\sigma}_p^2) < 0$ while $\hat{\sigma}_b^2$ is accurate in that $\mathrm{Bias}(\hat{\sigma}_b^2) = 0$. However, the variance of $\hat{\sigma}_b^2$ is larger than that of $\hat{\sigma}_p^2$, that is, $\mathrm{Var}(\hat{\sigma}_p^2) < \mathrm{Var}(\hat{\sigma}_b^2)$. Thus, elimination of the bias introduces additional variance (Figure 4.6). This is the bias-variance tradeoff. Moreover, in terms of the MSE, $\hat{\sigma}_p^2$ does better, that is, $\mathrm{MSE}(\hat{\sigma}_p^2) < \mathrm{MSE}(\hat{\sigma}_b^2)$. Therefore, neither estimator is unequivocally preferable over the other.

While it may be possible to bring the bias down to zero, it is not reasonable to expect that the variance of any sensible estimator is zero. So, the second natural criterion is to focus on the "most accurate" estimators and prefer one with a smaller variance. The estimator that has zero bias is called unbiased.

Definition 4.3 (Unbiasedness)**.** An estimator $\hat{\theta}$ of θ is *unbiased* if $\mathrm{Bias}(\hat{\theta}) = 0$.

Remark 4.1. Provided that $\hat{\theta}$ has a second moment, we can equivalently formulate unbiasedness as $\mathbb{E}[(\hat{\theta} - \theta)^2] \leq \mathbb{E}[(\hat{\theta} - b)^2]$ for every b (Theorem 2.5). This formulation allows us to generalize unbiasedness under other risk measures. For example, *median unbiasedness* of an estimator $\hat{\theta}$ can be defined by replacing the mean squared error with the mean absolute error, $\mathbb{E}[|\hat{\theta} - \theta|] \leq \mathbb{E}[|\hat{\theta} - b|]$ for every b (Theorem 2.6).

4.3. The Cramér–Rao Bound

What is the lowest possible variance attainable by an unbiased estimator? We answer this question for a specific class of models that is simple and smooth. In particular, we assume that the probability distribution of the data X is completely characterized by a finite-dimensional parameter θ. By this, I mean that once you know the value of θ, that uniquely pins down the probability distribution of X, so the pdf of X can be indexed by θ, i.e., $p_{n,\theta}(x_1, \ldots, x_n)$. Such a model is called the *parametric model*. Below are some examples of such models.

Example 4.9 (Normal location). Let $X = (X_1, \ldots, X_n)$ be a sample of i.i.d. k-dimensional random variables from $N(\theta, \Sigma)$ where Σ is known and symmetric positive definite. Since Σ is known, there is one-to-one relationship between the distribution and θ.

Example 4.10 (Bernoulli). Let $X = (X_1, \ldots, X_n)$ be a sample of i.i.d. Bernoulli random variables with $\theta = P(X_i = 1)$. Since a value of θ completely characterizes the probability of X, there is one-to-one relationship between the distribution and θ.

There are, of course, models that fall outside this class, such as the following.

Example 4.11. Let $X = (X_1, \ldots, X_n)$ be a sample of i.i.d. k-dimensional random variables from an unknown distribution. Let $\theta = \mathbb{E}[X_i]$ be the mean of X_i that we are interested in. There are many possible distributions that give rise to the same value of θ, so there is no one-to-one relationship between the distribution of X and θ.

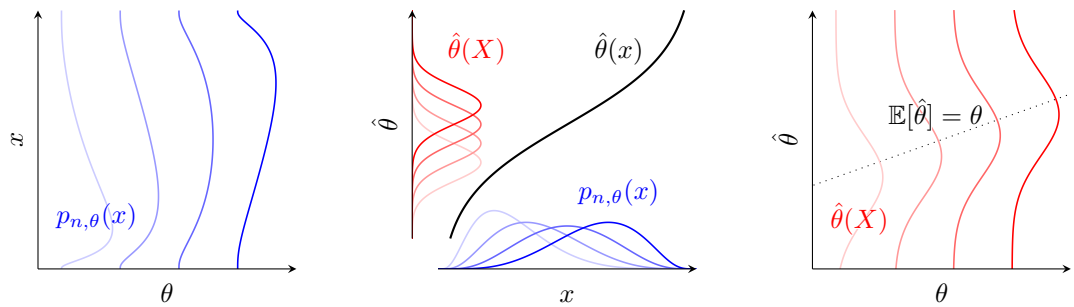
Going back to the parametric models, suppose $\hat{\theta}$ is an unbiased estimator for θ. Then, we have $\mathbb{E}[\hat{\theta}(X)] = \theta$ by definition. Letting $p_{n,\theta}$ be the pdf of the data $X = (X_1, \ldots, X_n)$, the LHS can be written as

$$\int \underbrace{\hat{\theta}(x)}_{\text{does not depend on } \theta} \overbrace{p_{n,\theta}(x)}^{\text{depends on } \theta} \, dx.$$

Since the function $\hat{\theta}(x)$ only depends on x and not on θ, the change in $\mathbb{E}[\hat{\theta}(X)]$ in response to the change in θ can only be induced through the change in $p_{n,\theta}(x)$ (Figure 4.7). If $\mathrm{Var}(\hat{\theta}(X))$ is very small, then $\hat{\theta}(x)$ is almost a constant function in x, and so would $\mathbb{E}[\hat{\theta}(X)]$ be in θ. For $\mathbb{E}[\hat{\theta}(X)]$ to move along with θ, therefore, the function $\hat{\theta}(x)$ must vary "enough" that $\mathbb{E}[\hat{\theta}(X)]$ can respond properly to the change in $p_{n,\theta}$. This argument suggests that the variance of an unbiased estimator cannot be too small.

Now, the rough sketch of the derivation of the bound goes as follows. Let θ be a univariate parameter and $\hat{\theta}$ unbiased. Then, we have $\mathbb{E}[\hat{\theta} - \theta] = 0$ for every θ. The resulting constraint on the changes in response to θ can be obtained by differentiating this identity with respect to θ,

$$\frac{d}{d\theta} \int (\hat{\theta}(x) - \theta) p_{n,\theta}(x) dx = \int (\hat{\theta}(x) - \theta) \dot{p}_{n,\theta}(x) - \int p_{n,\theta}(x) dx = 0.$$



(A) Distribution of X as a function of θ.

(B) Distribution of $\hat{\theta}$ as a function of $p_{n,\theta}$.

(C) Distribution of $\hat{\theta}$ as a function of θ.

FIGURE 4.7. Since $\hat{\theta}$ does not depend on θ, $\mathbb{E}[\hat{\theta}]$ moves only in response to the change in $p_{n,\theta}$. For $\hat{\theta}$ to be unbiased, $\hat{\theta}(x)$ cannot be too flat, which gives rise to a lower bound on $\mathrm{Var}(\hat{\theta})$.

Since the second integral is 1, this yields

$$1 = \int (\hat{\theta}(x) - \theta) \cdot \frac{\dot{p}_{n,\theta}(x)}{p_{n,\theta}(x)} \cdot p_{n,\theta}(x) dx \leq \sqrt{\mathrm{Var}(\hat{\theta}(X)) \mathbb{E}\left[\frac{\dot{p}_{n,\theta}(X)^2}{p_{n,\theta}(X)^2}\right]}$$

by the Cauchy–Schwarz inequality. Rearranging this gives a desired bound

$$\mathrm{Var}(\hat{\theta}(X)) \geq \mathbb{E}\left[\frac{\dot{p}_{n,\theta}(X)^2}{p_{n,\theta}(X)^2}\right]^{-1}.$$

The ratio $\dot{p}_{n,\theta}/p_{n,\theta}$ represents the rate of change of $p_{n,\theta}$ in response to θ. If $p_{n,\theta}$ changes very sensitively to a minuscule change in θ, then there is not much need for $\hat{\theta}$ to be sensitive to x, and hence its variance can be smaller. Conversely, if $p_{n,\theta}$ is not so responsive to the change in θ, then $\hat{\theta}$ must in turn be responsive enough to x in order to have $\mathbb{E}[\hat{\theta}]$ chase after θ. Thus, it makes sense that the bound is reciprocal to the "size" of $\dot{p}_{n,\theta}/p_{n,\theta}$. In a sense, the size (second moment) of $\dot{p}_{n,\theta}/p_{n,\theta}$ can be viewed as the *information* about θ contained in the model $p_{n,\theta}$; the larger it is, the more precise the estimator can possibly be. This argument generalizes to the parametric models that are "smooth" in the following sense.

Let $X = (X_1, \ldots, X_n)$ be a sample of n observations. The sample X has a joint pdf at every $x = (x_1, \ldots, x_n)$ indexed by $\theta \in \Theta \subset \mathbb{R}^k$ and denoted by $p_{n,\theta}(x)$. The density is assumed to be positive on a common support, that is, if $p_{n,\theta}(x) > 0$ for some θ at some x, then we have $p_{n,\theta'}(x) > 0$ for every $\theta' \in \Theta$ at this x. The density function is called the *likelihood function* when it is seen primarily as a function of θ. Since there is a one-to-one correspondence between the probability distribution and θ, we denote by $\mathbb{E}_\theta[f(X)]$ the expectation of $f(X)$ when X follows $p_{n,\theta}$. The *log likelihood function* $\ell_{n,\theta}(x) := \log p_{n,\theta}(x)$ is assumed to be differentiable with respect to θ, and its derivative $\dot{\ell}_{n,\theta}(x) = \frac{\partial}{\partial\theta}\ell_{n,\theta}(x) = \dot{p}_{n,\theta}(x)/p_{n,\theta}(x)$, called the *score function*, is a $k \times 1$ vector-valued function continuous in both θ and x. The *Fisher information matrix* $I_{n,\theta} := \mathbb{E}_\theta[\dot{\ell}_{n,\theta}(X)\dot{\ell}_{n,\theta}(X)']$ is assumed to be invertible.

Theorem 4.6 (Bartlett identities)**.** *Let $p_{n,\theta}$ be smooth as described above. Then, the expectation of the score is zero, i.e., $\mathbb{E}_\theta[\dot{\ell}_{n,\theta}(X)] = 0$. This means that the Fisher information is the variance of the score. If, moreover, $\ell_{n,\theta}$ is twice differentiable with the second derivative continuous in both θ and x, then we have another way to compute the Fisher information, $I_{n,\theta} = -\mathbb{E}_\theta[\ddot{\ell}_{n,\theta}(X)]$.*

Proof. Since $\int p_{n,\theta}dx = 1$ and $\dot{p}_{n,\theta}$ is continuous in θ and x, we have

$$\mathbb{E}_\theta[\dot{\ell}_{n,\theta}(X)] = \int \dot{\ell}_{n,\theta} p_{n,\theta} dx = \int \frac{\partial}{\partial\theta} p_{n,\theta} dx = \frac{d}{d\theta}\int p_{n,\theta}dx = \frac{d}{d\theta}1 = 0,$$

which is the first identity. If $\ell_{n,\theta}$ is twice differentiable,

$$\frac{\partial^2}{\partial\theta\partial\theta'}\ell_{n,\theta} = \frac{\partial}{\partial\theta'}\frac{\dot{p}_{n,\theta}}{p_{n,\theta}} = \frac{\ddot{p}_{n,\theta}}{p_{n,\theta}} - \frac{\dot{p}_{n,\theta}\dot{p}'_{n,\theta}}{p^2_{n,\theta}} = \frac{\ddot{p}_{n,\theta}}{p_{n,\theta}} - \dot{\ell}_{n,\theta}\dot{\ell}'_{n,\theta}.$$

If $\ddot{p}_{n,\theta}$ is continuous in θ and x,

$$\mathbb{E}_\theta\left[\frac{\ddot{p}_{n,\theta}}{p_{n,\theta}}\right] = \int \ddot{p}_{n,\theta}dx = \frac{d^2}{d\theta d\theta'}\int p_{n,\theta}dx = 0.$$

These yield the second identity. ■

Note that the above calculus holds just as well for a single observation, i.e., when $n = 1$. Then, if X is a sample of i.i.d. random variables, the score and information of the sample can be deduced from those of each observation. To distinguish the notation for a single observation and for the sample, denote by p_θ and $\ell_\theta = \log p_\theta$ the pdf and the log likelihood of each observation. Since $p_{n,\theta}(x) = p_\theta(x_1)\cdots p_\theta(x_n)$, we have

$$\ell_{n,\theta}(x) = \sum_{i=1}^n \ell_\theta(x_i), \qquad \dot{\ell}_{n,\theta}(x) = \sum_{i=1}^n \dot{\ell}_\theta(x_i),$$

$$I_{n,\theta} = \mathbb{E}_\theta\left[\sum_{i=1}^n\sum_{j=1}^n \dot{\ell}_\theta(X_i)\dot{\ell}_\theta(X_j)'\right] = \mathbb{E}_\theta\left[\sum_{i=1}^n \dot{\ell}_\theta(X_i)\dot{\ell}_\theta(X_i)'\right] = n\mathbb{E}_\theta[\dot{\ell}_\theta(X_i)\dot{\ell}_\theta(X_i)'] = nI_\theta.$$

Example 4.9 (Normal location, continued). Recall that $X = (X_1, \ldots, X_n)$ is a sample of i.i.d. normal random variables, $N(\theta, \Sigma)$ where Σ is known. Then, the pdf of one observation is given by $p_\theta(x) = (2\pi)^{-k/2}\det(\Sigma)^{-1/2}\exp(-\frac{1}{2}(x-\theta)'\Sigma^{-1}(x-\theta))$. Therefore, one can calculate

$$\dot{\ell}_\theta(x) = \Sigma^{-1}(x-\theta), \qquad \ddot{\ell}_\theta(x) = -\Sigma^{-1}, \qquad I_\theta = \Sigma^{-1}.$$

So, the Fisher information of the entire data X is $I_{n,\theta} = n\Sigma^{-1}$.

For the Bernoulli example, the pdf does not exist since the cdf is not differentiable. However, this framework can still be applied with a simple (but advanced) fix given in Section 2.A.

Example 4.10 (Bernoulli, continued). Recall that $X = (X_1, \ldots, X_n)$ is a sample of i.i.d. Bernoulli random variables with $\theta = P(X_i = 1)$. Let λ be a measure that has

unit masses at 0 and 1. Then, the density of a single observation with respect to λ is given as a probability mass function $p_\theta(x) = \theta^x(1-\theta)^{1-x}$. Therefore,

$$\dot{\ell}_\theta(x) = \frac{x}{\theta} - \frac{1-x}{1-\theta}, \qquad \ddot{\ell}_\theta(x) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}, \qquad I_\theta = \frac{1}{\theta(1-\theta)},$$

and the Fisher information of the entire data X is given by $I_{n,\theta} = \frac{n}{\theta(1-\theta)}$.

Now we state the general results. Just as derived in the rough sketch, the inverse of the Fisher information matrix gives the lower bound for the variance of an unbiased estimator.

Theorem 4.7 (Cramér–Rao bound)**.** *Let $X = (X_1, \ldots, X_n)$ be a sample from a smooth parametric model $p_{n,\theta}$. Let $\hat{\theta}(X)$ be an estimator of θ such that $\frac{d}{d\theta'}\mathbb{E}_\theta[\hat{\theta}(X)] = \int \hat{\theta}(x)\frac{\partial p_{n,\theta}}{\partial \theta'}dx$ holds (that is, differentiation under the integral sign). Then, we have $\mathrm{Var}(\hat{\theta}(X)) \geq (\frac{d}{d\theta'}\mathbb{E}_\theta[\hat{\theta}(X)])I_{n,\theta}^{-1}(\frac{d}{d\theta'}\mathbb{E}_\theta[\hat{\theta}(X)])'$. If $\hat{\theta}$ is unbiased, then $\mathrm{Var}(\hat{\theta}(X)) \geq I_{n,\theta}^{-1}$.*

Proof. By the Cauchy–Schwarz inequality for matrices [Tri99],

$$\mathrm{Var}(\hat{\theta}(X)) \geq \mathrm{Cov}(\hat{\theta}(X), \dot{\ell}_{n,\theta}(X))\,\mathrm{Var}(\dot{\ell}_{n,\theta}(X))^{-1}\,\mathrm{Cov}(\dot{\ell}_{n,\theta}(X), \hat{\theta}(X)).$$

Since $\mathbb{E}_\theta[\dot{\ell}_{n,\theta}(X)] = 0$ (Theorem 4.6), we have $\mathrm{Var}(\dot{\ell}_{n,\theta}(X)) = I_{n,\theta}$ and

$$\mathrm{Cov}(\hat{\theta}(x), \dot{\ell}_\theta(X)) = \int \hat{\theta}(x)\dot{\ell}_{n,\theta}(x)'p_{n,\theta}(x)dx = \int \hat{\theta}(x)\dot{p}'_{n,\theta}dx = \frac{d}{d\theta'}\mathbb{E}_\theta[\hat{\theta}(X)].$$

If $\hat{\theta}$ is unbiased, then $\frac{d}{d\theta'}\mathbb{E}_\theta[\hat{\theta}(X)] = \frac{d}{d\theta'}\theta$ is an identity matrix. ■

Note that Theorem 4.7 does not prove that an estimator attaining this bound exists.[3] But when it does, we know that it is the best estimator in terms of the variance. We call such an estimator efficient.

Definition 4.4 (Efficiency)**.** An unbiased estimator $\hat{\theta}$ of θ is *efficient* if it attains the Cramér–Rao bound, i.e., $\mathrm{Var}(\hat{\theta}) = I_{n,\theta}^{-1}$.

Example 4.9 (Normal location, continued). First, the sample average $\bar{X}_n$ is unbiased since $\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_i \mathbb{E}[X_i] = \theta$. Second, the variance of $\bar{X}_n$ is $\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_i \mathrm{Var}(X_i) = \Sigma/n$ by independence. This is equal to the Cramér–Rao bound, $I_{n,\theta}^{-1} = \Sigma/n$. Therefore, $\hat{\theta} = \bar{X}_n$ is an efficient estimator for θ.

Example 4.10 (Bernoulli, continued). Likewise, the sample average $\hat{\theta} = \bar{X}_n$ is an efficient estimator for θ.

When some assumptions of Theorem 4.7 are violated, there may be an unbiased estimator that attains a smaller variance than the Cramér–Rao bound.

Example 4.12 (Counterexample). Let $X = (X_1, \ldots, X_n)$ be i.i.d. observations from $U[0, \theta]$. Then $p_\theta(x) = \theta^{-1}\mathbb{1}\{0 < x < \theta\}$, but this density does not satisfy the common support condition. We can still calculate $\dot{\ell}_\theta(x) = -\theta^{-1}$ and $\ddot{\ell}_\theta(x) = \theta^{-2}$ for

[3] We revisit this point in Section 6.2.

$0 < x < \theta$, but then $\mathbb{E}[\dot{\ell}_\theta(X)] \neq 0$ and $I_\theta = \theta^{-2} \neq -\mathbb{E}_\theta[\ddot{\ell}_\theta(X)] = -\theta^{-2}$, so even the Bartlett identities (Theorem 4.6) fail. Also, since $\ell_{n,\theta}(x) = -n \log \theta$, we have that $I_{n,\theta} = n^2\theta^{-2} = n^2 I_\theta$, not $I_{n,\theta} = nI_\theta$. Now, consider the estimator $\hat{\theta} = \frac{n+1}{n} X_{(n)}$, where $X_{(n)}$ is the maximum observation. Then, $\hat{\theta}$ is unbiased and has a variance smaller than the Cramér–Rao bound. To see this, observe that

$$\begin{aligned} F_{X_{(n)}}(x) &= P(\max\{X_1, \ldots, X_n\} \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) \\ &= P(X_1 \leq x) \cdots P(X_n \leq x) = F_X(x)^n = x^n/\theta^n \end{aligned}$$

for $0 < x < \theta$, and hence $p_{X_{(n)}}(x) = nx^{n-1}/\theta^n$. With this, we can calculate

$$\mathbb{E}_\theta[X_{(n)}] = \int_0^\theta n\theta^{-n} x^n dx = \left[n\theta^{-n} \frac{x^{n+1}}{n+1} \right]_0^\theta = \frac{n\theta}{n+1}, \qquad \mathbb{E}_\theta[X_{(n)}^2] = \frac{n\theta^2}{n+2}.$$

These imply $\mathbb{E}_\theta[\hat{\theta}] = \theta$ and $\mathrm{Var}(X_{(n)}) = \frac{n}{(n+1)^2(n+2)}\theta^2$. So,

$$\mathrm{Var}(\hat{\theta}) = \frac{(n+1)^2}{n^2} \mathrm{Var}(X_{(n)}) = \frac{\theta^2}{n(n+2)} < I_{n,\theta}^{-1} = \frac{\theta^2}{n^2},$$

violating the Cramér–Rao bound.

4.4. Asymptotic Properties of Estimators

The efficiency of the estimator is a very appealing property, and yet it is not always useful for several reasons. First, not all parameters admit unbiased estimators. Second, for many reasonable estimators, we cannot even compute the bias analytically. Third, we do not want to restrict our attention only to the parametric models. Therefore, it makes sense to develop some criteria that are weaker but easier to achieve. For this, asymptotic criteria are the way to go. As we have seen in Chapter 3, the asymptotic analysis is a type of approximation, and as such, it makes calculation (and verification of the criteria) a lot easier.

The first desirable asymptotic property of an estimator is consistency, which requires that it converge in probability to the parameter of interest. For example, the sample average is a consistent estimator for the mean by the LLN (Theorem 3.4).

Definition 4.5 (Consistency)**.** An estimator $\hat{\theta}$ of θ is *consistent* if $\hat{\theta} \to^p \theta$.

The next desirable property is that the estimator converges in distribution to a normal distribution. Again, by the CLT (Theorem 3.5), the sample average is an asymptotically normal estimator for the mean.

Definition 4.6 (Asymptotic normality)**.** An estimator $\hat{\theta}$ of θ is $\sqrt{n}$*-regular* if $\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow L$ for some fixed distribution L that does not depend on θ. An estimator $\hat{\theta}$ is $\sqrt{n}$*-asymptotically normal* if it is $\sqrt{n}$-regular and $L = N(0, \Sigma)$ for some Σ.

As discussed in Section 3.2.2, note that the variance of $\hat{\theta}$ is Σ/n, not Σ, but the formal statement multiplies $\hat{\theta}$ by $\sqrt{n}$ to have a limit variance that is independent of n. To emphasize the fact that Σ is the "scaled" variance, Σ is sometimes called the *asymptotic variance*.

EXERCISE 4.4. Show that a $\sqrt{n}$-asymptotically normal estimator is consistent.

The criteria we developed in Section 4.2 can be translated to asymptotic versions, namely, an estimator that minimizes the MSE and an unbiased estimator that minimizes the variance.[4] While we do not elaborate it here, these two criteria, interestingly, lead to the same estimator, so there is often one unique estimator preferred in the asymptotic sense. This optimality criterion can be simply summarized by the notion of *parametric efficiency*.

Definition 4.7 (Parametric efficiency)**.** Let $\{P_\theta\}$ be a parametric model with a nonsingular Fisher information matrix I_θ that will be defined in Section 4.3. An estimator $\hat{\theta}$ of θ is *asymptotically parametrically efficient* if $\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow N(0, I_\theta^{-1})$.

Remark 4.2. There is an extension of parametric efficiency to semiparametric models, called *semiparametric efficiency* [**vdV98**, Chapter 25].

In econometrics, a weaker version of efficiency is also used, which only requires that an estimator have the minimum asymptotic variance among some restricted class of estimators (e.g., efficient GMM [**Hay00**, Proposition 3.5]).

4.5. Three Types of Statistical Modeling

If calculating descriptive statistics is about knowing the data, carrying out statistical analysis is about knowing more than the data. The "more" part hinges on the additional assumption we are willing to buy, and it is informally referred to as the "model." There are various assumptions we can make in various situations, and it is helpful to classify them into three categories: parametric models, semiparametric models, and nonparametric models.

In practical terms, *parametric models can be described as follows.

(1) *Parametric models.* The model with a complete description of the probability structure, including the shape of the distributions. This requires strong assumptions and tends to yield precise estimators.
(2) *Semiparametric models.* The model that is half-specified. There is a parameter of interest, while some other aspects of the probability structure are left unspecified. This requires mild assumptions and tends to yield fair estimators.
(3) *Nonparametric models.* The model that is hardly specified. There is a parameter of interest, but it is much like a descriptive statistic. This requires weak assumptions and may yield not-so-precise estimators.

In mathematical terms, the descriptions are a bit involved. Recall that a parameter is defined as a function of the probability distribution $P \mapsto \theta(P)$.

(1) *Parametric models.* There is a finite-dimensional parameter θ that fully characterizes the probability distribution, i.e., there exists a one-to-one correspondence between the parameter and the probability distribution, $\theta \leftrightarrow P$.

[4]Roughly, the minimum MSE estimator corresponds to the *local asymptotic minimax theorem* and the minimum variance unbiased estimator to the *convolution theorem* [**vdV98**, Chapter 8].

This implies that the set of probability distributions under consideration does not span all of the probability distributions, and that the parameter map is invertible, $\theta \mapsto P(\theta)$. The parameter *of interest* may be a subset of θ, but this is usually not an issue.

(2) *Semiparametric models.* There is a finite-dimensional parameter θ and an infinite-dimensional parameter η that collectively characterizes the probability distribution, i.e., $(\theta, \eta) \leftrightarrow P$, and the set of probability distributions does not span all of the probability distributions. This means that the map $P \mapsto \theta$ is not invertible, which makes the theory of semiparametric models quite complicated. The parameter of interest is usually (a subset of) θ.

(3) *Nonparametric models.* There is an infinite-dimensional parameter η that characterizes the probability distribution, i.e., $\eta \leftrightarrow P$, and the set of probability distributions may or may not span all of the probability distributions. Sometimes η is itself taken to be P, and η is the parameter of interest.

Note that they are not mutually exclusive nor nested. For example, if you add a finite-dimensional parameter to a nonparametric model, it becomes a semiparametric model and more general than the original nonparametric model.

Example 4.13 (Normal location model). Let $\Theta = \mathbb{R}^k$ and $\mathcal{P} = \{N(\mu, \Sigma) : \mu \in \Theta\}$ be the set of all univariate normal distributions with some fixed known variance Σ. The parameter μ is the mean of the normal distribution, and there is a one-to-one correspondence between μ and P. This is the *normal location model*. If we also treat Σ as an unknown parameter, it is called the *normal location-scale model*. These are examples of the parametric model.

Example 4.14 (Single-index model). Suppose there is a triplet $(Y_i, X_{i1}, X_{i2}) \sim P$. The model $\mathcal{P}$ contains all distributions with finite variances and $\mathbb{E}[Y_i \mid X_i = x] = g(x'\beta)$ for some $\beta \in \mathbb{R}^2$ and $g : \mathbb{R} \to \mathbb{R}$ that is smooth. This means that the dependence of Y on $X = (X_1, X_2)'$ only comes from the linear term $X_i'\beta = \beta_1 X_{i1} + \beta_2 X_{i2}$ (called the index), but the dependence of Y_i on this linear term is left unspecified. The parameter of interest may be both of β and the function g. This is known as the *single-index model* and is a special case of the *projection pursuit regression*. This is an example of the semiparametric model.

Example 4.15 (Descriptive statistics). Suppose there is a univariate variable $X_i \sim P$ and take the model $\mathcal{P}$ to be the set of all univariate distributions. Then the cdf of X_i is well defined for every $P \in \mathcal{P}$ and can be the parameter of interest. Or, the median of X_i is also well defined and can be the parameter of interest. This is an example of the nonparametric model. If we restrict $\mathcal{P}$ to span all univariate distributions with finite variances, then the mean of X_i can be the parameter of interest. This can still be regarded as a nonparametric model.

Example 4.16 (Nonparametric regression). Suppose there is a pair $(Y_i, X_i) \sim P$. Let the model $\mathcal{P}$ span all bivariate distributions with finite variances and smooth conditional expectation $\mathbb{E}[Y_i \mid X_i = x]$. This can equivalently be written as $Y_i = g(X_i) + \varepsilon$ where g is smooth and $\mathbb{E}[\varepsilon \mid X_i] = 0$. The parameter of interest is the

function g. This is known as the *nonparametric regression* and is an example of the nonparametric model. We may allow X_i to be multidimensional.

CHAPTER 5

Principles of Statistical Inference

If you torture the data enough, nature will always confess.

HOW SHOULD ECONOMISTS CHOOSE? RONALD H. COASE, 1982

Given an estimator of the parameter of interest, many decision problems can be solved by simply plugging the estimates into the unknown parameters. For example, an optimal portfolio may be derived by maximizing the utility function whose parameters are replaced with their estimates. In some cases, however, the decision problems cannot be disentangled from the consideration of how close the estimates are to the true parameter values.

Consider a media calling an election from the results of the exit poll. With the estimate of the excess vote of 2%, can we conclude that the first runner is winning the election? Here, the question is not about the value of 2%, but rather about whether the 2% is a sufficient indicator that the true excess vote is positive. In another example, consider deploying the Patriot missiles for national defense. Let us say that the goal of this defense system is to make the probability of successful interception (taking down incoming ballistic missiles in the sky) as high as 99.9%. If the data from several test shots give the exact estimate of 99.9%, can we trust that we are protected with high certainty? More likely, we may be worried about the possibility that the test shots were merely "good" draws by chance, and the probability of actual interception may not be as high as expected. How can we guard against such a worst case scenario?

Although the true parameter values are never known, statistics offers ways to make reasonable decisions under these circumstances. Consideration of the first question leads to *hypothesis testing* and the concept of the *p-value*; consideration of the second to the *confidence interval*.

5.1. Extracting a Simple Experiment

To focus on the essential aspects, statistical inference is often discussed in a vastly simplified setup. To illustrate the simplification process, take the problem of estimating the mean θ of a univariate variable X with an unknown finite variance σ^2. We have the sample of n i.i.d. observations $X_1, \ldots, X_n$, and we estimate θ by the sample mean $\hat{\theta} = \bar{X}_n$. By the CLT, we know that $\hat{\theta}$ converges in distribution to $N(\theta, \sigma^2/n)$.

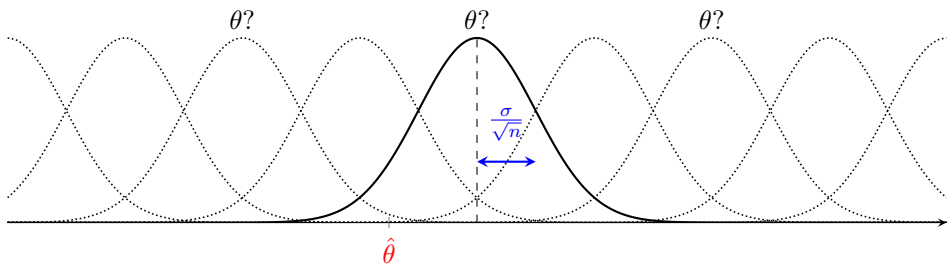


FIGURE 5.1. Normal location model. We draw one and only one observation $\hat{\theta}$ from a normal distribution $N(\theta, \sigma^2/n)$ where θ is unknown and σ^2/n is known.

Although we do not know σ^2, it can be regarded as yet another estimation problem and we can estimate it by the sample variance $\hat{\sigma}^2 = \frac{1}{n-1}\sum(X_i - \bar{X}_n)^2$. Moreover, $\hat{\sigma}^2$ converges in probability to σ^2 by the LLN,[1] so the difference between $\mathrm{Var}(\hat{\theta}) = \sigma^2/n$ and $\widehat{\mathrm{Var}}(\hat{\theta}) = \hat{\sigma}^2/n$ vanishes at a rate *faster* than $1/n$, the rate at which $\mathrm{Var}(\hat{\theta})$ shrinks to zero. That is,

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \underbrace{\mathrm{Var}(\hat{\theta})}_{O(1/n)} + \underbrace{\widehat{\mathrm{Var}}(\hat{\theta}) - \mathrm{Var}(\hat{\theta})}_{o_P(1/n)}.$$

In large samples, therefore, we can ignore the estimation error of $\widehat{\mathrm{Var}}(\hat{\theta})$ and pretend that we *know* $\mathrm{Var}(\hat{\theta})$. Thus, ultimately we have the following simple situation: the population model $N(\theta, \sigma^2/n)$ where we know σ^2/n, and we have *one and only one* observation $\hat{\theta}$ therefrom (Figure 5.1).

This reduction takes place in many statistical problems. At the beginning, we have a possibly very complicated population model $\mathcal{P}$ and a possibly quite high-dimensional dataset X (Figure 4.1). Then, we choose an estimation method and estimate θ by $\hat{\theta}$. In many cases, $\hat{\theta}$ converges in distribution to a normal distribution centered at θ with variance $\mathrm{Var}(\hat{\theta})$, which we can estimate by some consistent estimator $\widehat{\mathrm{Var}}(\hat{\theta})$. By the same logic as above, we can ignore the estimation error associated with the variance estimator (when the sample size is large), so we assume that we know $\mathrm{Var}(\hat{\theta})$. Thus, we have in our hands the vastly simplified "normal location model" (or "normal experiment") with one observation (Figure 5.2). Of course, there can also be other cases such as $\hat{\theta}$ converges to a non-normal distribution, the mean of $\hat{\theta}$ is not θ, or the analysis does not involve estimation of θ; however, some form of simplification is almost always possible and the extracted model is called the *experiment*.

EXERCISE 5.1 (Variance estimation). Let $X_1, \ldots, X_n$ be i.i.d. normal random variables from $N(\mu, \sigma^2)$. Let $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Show that

$$P\Big(\hat{\mu} \in \Big[\mu - 2\tfrac{\hat{\sigma}}{\sqrt{n}}, \mu + 2\tfrac{\hat{\sigma}}{\sqrt{n}}\Big]\Big) - P\Big(\hat{\mu} \in \Big[\mu - 2\tfrac{\sigma}{\sqrt{n}}, \mu + 2\tfrac{\sigma}{\sqrt{n}}\Big]\Big) \longrightarrow 0.$$

[1] If X has a fourth moment, the rate at which $\hat{\sigma}^2$ converges to σ^2 is $1/\sqrt{n}$ by the CLT. Otherwise, it can be slower.

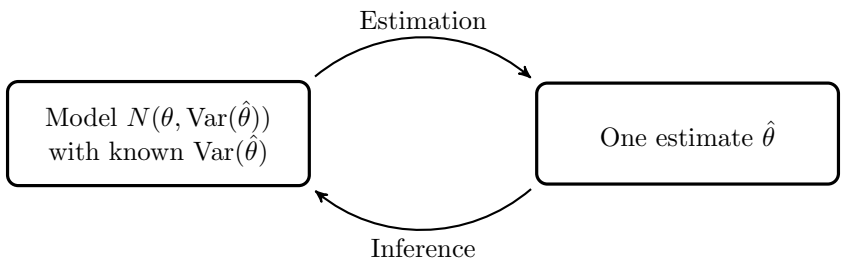


FIGURE 5.2. Normal experiment after estimation.

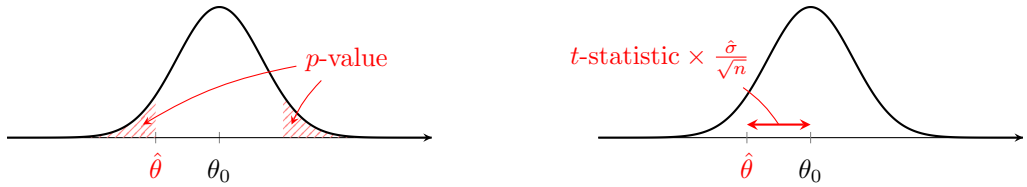
5.2. Hypothesis Testing

Let us continue on the example of calling an election. When calling an election, what we are worried about is that the actual winner is different. Assume that there are only two runners and let θ be the excess vote of candidate A against candidate B in the actual ballots. After the poll closes, there is a deterministic value of θ which is unknown until all the ballots are counted. Therefore, there is no "probability" that candidate A wins; the winner is already determined, and we simply do not know the result yet. So, how do we assess the likeliness of candidate A being the winner, given an estimate of $\hat{\theta} = 2\%$ and $\widehat{\mathrm{Var}}(\hat{\theta}) = (1.5\%)^2$ from the exit poll?

For this, it is natural to use the fact that $\hat{\theta}$ is observable and random. Since there is a probability of $\hat{\theta}$ realizing in a certain range, we can use it to measure the likeliness of a specific value of θ. For example, are we worried that the true excess vote may be $\theta = -4\%$ and candidate B wins? Not really, because if that is the case, then obtaining an estimate of 2% is extremely unlikely. That is, $\hat{\theta}$ distributes according to a normal distribution centered at -4% with the standard deviation of 1.5%, so 2% is "four-sigma away" from the mean. But how about $\theta = 0\%$ (no winner)? This time, we may very much be worried about the possibility of this scenario, since $\hat{\theta}$ realizing at 2% is a very conceivable event even if the true excess vote is 0%. Therefore, we come to the conclusion that we cannot call the election yet and should continue to collect responses from the exit poll.

Formalization of this reasoning is called hypothesis testing. Mathematically, a *hypothesis* is the assertion that θ is in some subset Θ_0 of Θ, denoted by $H : \theta \in \Theta_0$. It often takes the form of an equality or inequality, e.g., $H : \theta = 0$ or $H : \theta \geq 0$. A hypothesis that pins down the parameter uniquely (i.e., Θ_0 being a singleton) is called the *point* (or *simple*) *hypothesis*, and a hypothesis that allows multiple values of the parameter is called the *composite hypothesis*. The first step of hypothesis testing is to formulate the hypothesis that *we wish to turn down*. In this example, the media wants to turn down the hypothesis $H : \theta = 0$, since then its rejection leads to either $\theta > 0$ or $\theta < 0$, i.e., there is a winner. This hypothesis we wish to reject is called the *null hypothesis* and is denoted by H_0. Thus, the expression $H_0 : \theta = 0$ means that we wish to reject the hypothesis that θ is equal to zero.

Let us look at some more examples to solidify the concept. When a labor economist wants to assess the causal effect θ of a job training program on the chance of



(A) The p-value is the area under the hypothesized pdf that is further away from the hypothesis than $\hat{\theta}$. It follows $U[0,1]$ under the null (Exercise 5.4).

(B) The t-statistic is the measure of how many standard errors away $\hat{\theta}$ is from the hypothesized value. It follows $N(0,1)$ under the null.

FIGURE 5.3. Two measures of extremeness of the estimate $\hat{\theta}$ under the null hypothesis $H_0 : \theta = \theta_0$.

employment, her null hypothesis is $H_0 : \theta \leq 0$, the rejection of which implies that the job training program improves the chance of employment. When an auditor wants to guarantee that a client's compliance rate r is higher than 99%, her null hypothesis is $H_0 : r \leq 99\%$. The null hypothesis may also take a more complicated form. When a group of physicists declared the existence of a Higgs boson, their reasoning was by statistical inference; they plotted the distribution of the invariant mass of the diphoton events and compared it against the theoretical distribution under the assumption that there is no Higgs boson. In this case, the null hypothesis is that a function (the distribution) takes a specific shape. Finally, note that even when the ultimate interest is an inequality, practitioners often casually test the equality hypothesis; e.g., the labor economist may test $H_0 : \theta = 0$.

When we have a point null hypothesis, $H_0 : \theta = \theta_0$, it pins down the candidate distribution uniquely. Under this distribution, the probability that $\hat{\theta}$ would have realized as extreme as it is observed is called the *p-value* for H_0. Visually, it corresponds to the area under the normal pdf that is further than $\hat{\theta}$ from the mean in either tails (Figure 5.3A). When the null hypothesis is composite, the p-value is calculated as the maximum probability that $\hat{\theta}$ could have realized further away from Θ_0 as observed, where maximum is taken over all possible distributions in Θ_0 [**Was04**, Theorem 10.12].

In the normal location model, how "extreme" $\hat{\theta}$ has realized can also be measured by how many standard errors away $\hat{\theta}$ is from the hypothesized value θ_0. This is called the *t-statistic* (Figure 5.3B), that is,[2]

$$t := \frac{\hat{\theta} - \theta_0}{\hat{\sigma}/\sqrt{n}}.$$

As $\hat{\theta}$ follows the normal distribution with mean θ_0 and variance σ^2/n, the t-statistic follows the standard normal distribution under $H_0 : \theta = \theta_0$. Therefore, it is related

[2]In finite-sample testing theory, it is called the t-statistic when the denominator is estimated (i.e., the standard error) and the z-statistic when the denominator is known (i.e., the standard deviation). We do not make this distinction.

to the p-value through $p = 2[1 - \Phi(|t|)]$, where Φ is the cdf of the standard normal distribution.

The next step of hypothesis testing is to choose the size and draw a conclusion. The *size* is a number of our choice between 0 and 1, intuitively representing our tolerance toward rejecting a null hypothesis when it is correct ("type I error"); more precisely, it is the probability of rejection—or a bound thereof—when the null hypothesis is correct. Since our desired conclusion is the rejection of the null hypothesis, we want to be conservative in reaching that conclusion. Therefore, the size is usually chosen to be a small number, say 0.05. Then, we *reject* the null hypothesis if its p-value is less than the size, and *accept* it otherwise. Note that the acceptance of a null hypothesis is not a strong support in favor of the null; it merely indicates that the dataset was inconclusive for the hypothesis.[3] For this reason, some researchers prefer an alternative expression such as "fail to reject" or "retain" instead of "accept."

The more critical the consequence of falsely rejecting the null is, the smaller the size should be. Social science conventionally uses 0.05. The discovery of the Higgs boson was announced with a size smaller than 10^{-6}, since falsely dismissing the Higgs boson would substantially delay the progress of science. A firm may make marketing decisions with a larger size such as 0.1–0.25 as the consequence of ineffective marketing is usually only pecuniary.

When the sample size increases, the denominator of the t-statistic decreases. Meanwhile, if the null hypothesis is correct, $\hat{\theta}$ approaches θ_0 at the same rate, leaving the distribution of t unchanged. However, if the null hypothesis is incorrect, then $\hat{\theta}$ converges to a different value, so the numerator stays roughly constant and t is pushed away from zero. Therefore, the p-value of a wrong hypothesis gets smaller and smaller, making the hypothesis more and more likely to be rejected.

Exercise 5.2 (Null hypothesis). Suppose that a financial analyst wants to investigate if a hedge fund produces a positive alpha. What is her null hypothesis?

Exercise 5.3 (Trivial test). Suppose we have $\hat{\theta} \sim N(\theta, \sigma^2/n)$ where σ^2/n is known. Suppose also that we have $U \sim U[0,1]$ independent of $\hat{\theta}$. Then, rejecting when $U < 0.05$ and accepting when $U \geq 0.05$ gives a valid test of $H_0 : \theta = 0$ with size 5%, that is, the probability of rejection when the null hypothesis is correct is 5%. Why is this test "worse" than the test based on the p-value?

Exercise 5.4 (Distribution of the p-value). Show that the p-value for the hypothesis $H_0 : \theta = \theta_0$ in the univariate normal location model distributes according to a uniform distribution $U[0,1]$ under the null. *Hint: Recall Exercise 2.2.*

Exercise 5.5 (Distributions under alternatives). If the null hypothesis $H_0 : \theta = \theta_0$ is incorrect, how would the distributions of p and t change? Explain by words.

[3] This is because we did not control the probability of a "type II error," the false acceptance of the null when the null is actually wrong.

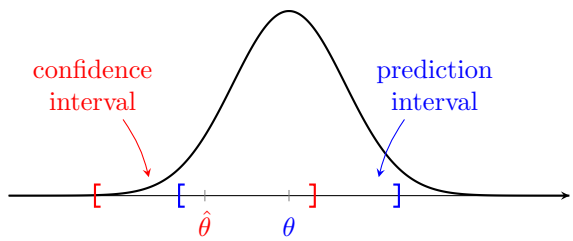


FIGURE 5.4. The confidence interval for θ is the flip side of the prediction interval for $\hat{\theta}$; θ is in the confidence interval if and only if $\hat{\theta}$ is in the prediction interval.

5.3. Confidence Intervals

In the example of national defense, there is no specific value of the interception probability that we wish to test. Rather, we want to guard against the possibility of drawing deceptively good estimates. This can be done by taking into account the "worst case" scenarios that could have given rise to the same estimate.

Suppose that the defense system consists of multiple Patriot missiles and that each missile has independent probability of interception denoted by θ. Suppose also that the test shots give an estimate of $\hat{\theta} = 70\%$ and $\widehat{\mathrm{Var}}(\hat{\theta}) = (4.6\%)^2$. If we believe the estimate 70% at its face value, then deploying 6 Patriot missiles is enough to guarantee a 99.9% interception probability, since $1 - (1 - 70\%)^6 > 99.9\%$. However, our estimate of 70% may be optimistic, and we want to know how conservative we should be. Should we worry that the true interception probability is 50%? No really, because then, obtaining an estimate of 70% is very unlikely; it is more than a "four-sigma" event. How about $\theta = 65\%$? For this, we should be worried about it since observing $\hat{\theta} = 70\%$ when the true probability is 65% is conceivable. This consideration gives rise to a range of plausible values of θ given an estimate $\hat{\theta}$.

Formalization of this is known as the confidence interval. Denote $\sigma^2/n = \mathrm{Var}(\hat{\theta})$ for simplicity. Given some θ, the range of plausible realizations of $\hat{\theta}$ is given by a prediction interval for $\hat{\theta}$. In particular, if $\hat{\theta}$ is normally distributed, $\hat{\theta}$ realizes within two standard deviations from the mean with 95% probability (Example 2.10), so the range $[\theta - 2\sigma/\sqrt{n}, \theta + 2\sigma/\sqrt{n}]$ gives the *95% prediction interval* for $\hat{\theta}$. In turn, we want the range of θ such that the observed $\hat{\theta}$ is in the prediction interval of every θ therein. This turns out to be easy since $\hat{\theta}$ is in $[\theta - 2\sigma/\sqrt{n}, \theta + 2\sigma/\sqrt{n}]$ if and only if θ is in $[\hat{\theta} - 2\sigma/\sqrt{n}, \hat{\theta} + 2\sigma/\sqrt{n}]$. This latter interval is called the *95% confidence interval* for θ (Figure 5.4). Mathematically,

$$P\Big(\underbrace{\hat{\theta}}_{\substack{\text{random}\\ \text{observable}}} \in \underbrace{\Big[\theta - 2\tfrac{\sigma}{\sqrt{n}}, \theta + 2\tfrac{\sigma}{\sqrt{n}}\Big]}_{\substack{\text{nonrandom}\\ \text{unobservable}}}\Big) = 95\% \iff P\Big(\underbrace{\theta}_{\substack{\text{nonrandom}\\ \text{unobservable}}} \in \underbrace{\Big[\hat{\theta} - 2\tfrac{\sigma}{\sqrt{n}}, \hat{\theta} + 2\tfrac{\sigma}{\sqrt{n}}\Big]}_{\substack{\text{random}\\ \text{observable}}}\Big) = 95\%.$$

It is important to understand that the randomness of the confidence interval comes from the randomness of $\hat{\theta}$. This means that once $\hat{\theta}$ realizes (that is, once you obtain an estimate using the dataset), there is no randomness involved in the confidence

interval. Therefore, when you see an interval of, e.g., $[-1, 1]$, it is not correct to think that there *is* a 95% chance that the true θ is in this range; instead, there *was* a 95% chance that this range could have realized and contained θ. Once you observe a solid interval, whether the true θ is within it is simply deterministic and unknown. Thus, while the confidence interval can be roughly thought of as giving the range of likely values for θ, this likeliness is given not as a probability, only as our "confidence."[4]

Back to the national defense example, we may pick the worst case parameter from the confidence interval to guard against the bad draw case. In particular, we may assume that the true intercept probability is as bad as $70\% - 2 \times 4.6\% = 60.8\%$. Then, deploying 8 Patriot missiles will make the intercept probability larger than 99.9%.

Remark 5.1. More generally, the $(1-\alpha)$-*confidence set* for θ is a random subset of Θ such that the probability that it contains true θ is at least $1-\alpha$. This is by no means unique, and sometimes a preference is given toward a narrower confidence set or a set with a desired shape (e.g., connected or rectangular).

5.4. Equivalence of the Two

The two concepts we have introduced, hypothesis testing and the confidence interval, are indeed only different sides of the same coin. In particular, the following statements are all equivalent.

- The hypothesis $\theta = \theta_0$ is rejected with size 5%.
- The p-value for the hypothesis $\theta = \theta_0$ is less than 5%.
- The t-statistic for the hypothesis $\theta = \theta_0$ is greater than 2 in magnitude.
- The 95% confidence interval does not contain θ_0.

Here, "2" is called the *critical value* and "95%" the *confidence level*. This equivalence is rooted in the fact that all of these statements are mere paraphrases of the following equation. Given an estimator $\hat{\theta} \sim N(\theta, \sigma^2/n)$, consider

$$P\left(|\hat{\theta} - \theta| > c\,\frac{\sigma}{\sqrt{n}}\right) = p.$$

There are three "free" parameters (θ, c, p), so we can solve for one by fixing two.

- If we fix $\theta = \theta_0$ and $c = 2$ and solve for p, we obtain the p-value for the hypothesis $\theta = \theta_0$ with size 5%.
- If we fix $\theta = \theta_0$ and $p = 0.05$ and solve for c, we obtain the critical value for the hypothesis $\theta = \theta_0$ with size p.
- If we fix $c = 2$ and $p = 0.05$ and solve for θ, we obtain the 95% confidence interval for θ.

This equation is quite handy when you recall the correct interpretation of the inferential concepts.

Remark 5.2. In some applications, direct construction of a confidence set can be harder than hypothesis testing. In such cases, a confidence set may be derived by

[4]The term "confidence" is likely attributed to Charles Sanders Peirce.

"inverting" a test, that is, a $(1-\alpha)$-confidence set can be formed as the collection of points at which the point null hypothesis is not rejected when tested with size α.

Remark 5.3. Note that, for statistical inference to be valid asymptotically, we do not need convergence of moments of $\hat{\theta}$ (Exercise 3.4); all we need is convergence in distribution to a normal and a consistent estimator of the asymptotic variance.

5.5. Testing Multivariate Hypotheses

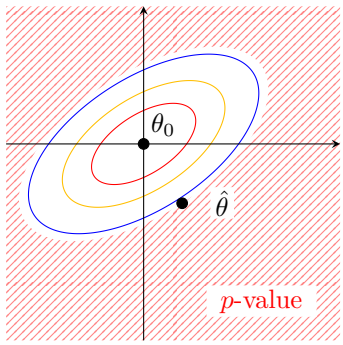
Suppose we have two parameters $\theta = (\theta_1, \theta_2)$ and want to test a hypothesis regarding both of θ_1 and θ_2. There are two important cases to discuss: a hypothesis of the type $\theta_1 = \theta_{0,1}$ *and* $\theta_2 = \theta_{0,2}$ and a hypothesis of the type $\theta_1 = \theta_{0,1}$ *or* $\theta_2 = \theta_{0,2}$. The *and* hypothesis is still a point hypothesis and is the subject of this section; the *or* hypothesis is composite and is the subject of multiple testing (Section 5.6). Since $H_0 : \theta_1 = \theta_{0,1}$ and $\theta_2 = \theta_{0,2}$ is a point hypothesis, we only need to decide how to measure the extremeness of a vector $\hat{\theta}$ from another θ_0 in order to generalize the statistical inference seen above.

In a canonical case, an estimator $\hat{\theta}$ of θ converges in distribution to a *joint* normal distribution $N(\theta, \Sigma/n)$ for a positive definite matrix Σ, and we also have a consistent estimator $\hat{\Sigma}$ for Σ. Note that the contour lines of the joint normal pdf are given as nested ellipses (Figure 5.5A). When hypothesis testing is concerned, the extremeness of $\hat{\theta}$ from θ_0 is usually measured by how far the ellipse to which $\hat{\theta}$ belongs is away from θ_0, which is the center of the distribution. The p-value is the volume under the pdf that is outside the ellipse of $\hat{\theta}$. This can be computed as follows. First, observe that the normalized vector $\sqrt{n}\Sigma^{-1/2}(\hat{\theta} - \theta_0)$ preserves the contours[5] and follows a multivariate standard normal distribution. Ergo, what is known as *Hotelling's T^2 statistic*, $T^2 = \|\sqrt{n}\hat{\Sigma}^{-1/2}(\hat{\theta} - \theta_0)\|^2$, follows a χ^2 distribution (Example 2.8). The p-value for H_0 is then given by the tail probability of a χ^2 distribution (with the degree of freedom equal to the number of parameters) above the observed T^2.

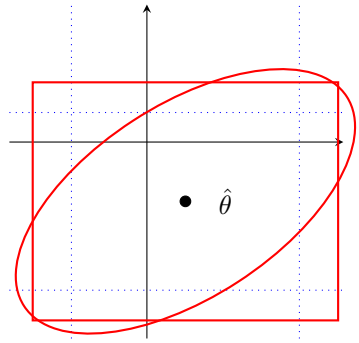
A confidence interval can also be derived with this ellipse-based extremeness measure, in which case it is an ellipse centered at $\hat{\theta}$. However, for ease of interpretation, a rectangle-based extremeness measure is also popular. Usually, the lengths of the sides are taken proportional to their marginal standard deviations, and we pick the rectangle centered at $\hat{\theta}$ whose coverage probability is as desired (Figure 5.5B). Another benefit of a rectangular confidence set is that each edge gives a valid marginal confidence interval, whilst being more conservative than the sharp univariate one.

As a minor note, univariate statistical inference can be viewed as bivariate statistical inference with an extremeness measure taking into account only one coordinate, e.g., a univariate confidence interval is a bivariate confidence "belt" that is infinitely long. In Figure 5.5B, the intersection of the confidence belts for the two parameters is shown as a blue dotted rectangle; note that this intersecting rectangle is narrower

[5]If two points belong to the same contour of the distribution of $\hat{\theta}$, they belong to the same contour of the distribution of $\sqrt{n}\Sigma^{-1/2}(\hat{\theta} - \theta_0)$, and vice versa.



(A) Joint hypothesis testing is usually done with an elliptic p-value (red shaded area) and an elliptic critical region.



(B) Joint confidence sets are usually a rectangle or an ellipse (red). The intersection of marginal confidence bands (blue) does not provide correct joint coverage.

FIGURE 5.5. Joint statistical inference for two parameters.

than the joint confidence rectangle, and hence fails to provide the intended coverage. The same observation also appeared in Figure 4.3.

EXERCISE 5.6. Find Σ such that the elliptic confidence set will be a strict subset of the rectangle confidence set at the same confidence level.

5.6. The Simultaneous Inference Problem and Multiple Testing

Statistical discovery is made on the basis of the rarity of our observation; if something extremely rare is observed, we take it as a counter-evidence to the postulated hypothesis and reject it in favor of the alternative hypothesis. Since the p-value distributes as $U[0,1]$ under the null (Exercise 5.4), a very small p-value must have come out with probability as low as the p-value itself had the hypothesis been true.[6] However, if we distort the distribution of the p-value and increase the chance of getting a small value, the rejection of the hypothesis does not constitute a legitimate discovery.

This can happen when we test many hypotheses and cherry-pick the ones to report. Suppose for simplicity that we test a correct null hypothesis many times independently. If we test twice, the probability that the minimum of the two p-values is less than 5% is 9.75%.[7] If we test thrice, it amounts to 14.26%, much less rare than 5%. Thus, even if the null hypothesis is correct, we can fabricate a rejection by repeating hypothesis testing and hiding the accepted ones; however, the rarity upon which the statistical discovery is based is lost. This is known as the *simultaneous inference problem*. We will revisit this issue in Section 7.5.

[6] More generally, the p-value has first-order stochastic dominance over $U[0,1]$ under the null [**LR05**, Lemma 3.3.1].

[7] A similar calculation as Example 3.5 shows that the cdf of the minimum of k independent standard uniform variables is $F(x) = \int_0^x k(1-x)^{k-1}dx$ for $0 \le x \le 1$.

A correct way to account for this problem is to create a huge hypothesis

$$H_0 : \theta_1 = \theta_{0,1} \quad \text{or} \quad \theta_2 = \theta_{0,2} \quad \text{or} \quad \theta_3 = \theta_{0,3} \quad \text{or} \quad \dots$$

and test it once. However, this turns out to be harder than it looks. A natural extremeness measure for this hypothesis would be the minimum of how far $\hat{\theta}_j$ is from $\theta_{0,j}$, that is, $\min_j |t_j|$ where t_j is the t-statistic for θ_j. If we reject H_0 when this minimum is sufficiently large, it is possible to test this hypothesis controlling the size at a desired level. However, there are at least two problems in this approach. First, if one of the equalities is violated (so one t-statistic follows an uncentered normal), the distribution of the minimum changes only so slightly, and we cannot expect the power of this test to be very high; this is especially pronounced when there are many parameters being examined. The alternative hypothesis for which this test has power is when a substantial portion of the equalities are violated. Second, in many applications, the joint distribution of $\hat{\theta}$ is not known, so we need to rely on some generic method that is more conservative than necessary. This worsens the problem of low power.

To cope with this issue, the notion of statistical discovery is loosened. The probability of rejection when even just one equality is violated is called the *familywise error rate (FWER)*. On the other hand, a weaker notion called the *false discovery rate (FDR)* is proposed and used. It aims to control the expected proportion of rejections, instead of the probability of one or more rejections. The key quantity used for this is called the *q-value* in analogy with the p-value. For more details on multiple testing, see [**LR05**, Chapter 9]. For a concrete use case of multiple testing in economics, see, for example, [**BDG+15**].

5.A. Finite-Sample Testing with Normality

The key to obtaining a normal-location-based inference was the convergence of our estimator to a normal distribution. This may provoke speculation that the same inferential procedure applies as long as our estimator is normal. For example, when each of X_i follows a normal distribution, the sample average $\bar{X}_n$ is exactly normal for every n, regardless of how small n is. However, such idea is not entirely justifiable for when the sample size is small, the randomness arising from $\hat{\sigma}^2$ cannot be ignored.

For $X_i \sim$ i.i.d. $N(\mu, \sigma^2)$ and $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, the exact distribution of the t-statistic $t = \frac{\bar{X}_n - \mu}{\hat{\sigma}/\sqrt{n}}$ is known as the *t-distribution with $n-1$ degrees of freedom*. This distribution of course converges to the standard normal distribution as the sample size grows, but the smaller the sample size, the wider the distribution, and hence the more conservative the inference need to be relative to the normal location model.

Similarly, when X_i is a $k \times 1$ i.i.d. random vector that follows $N(\mu, \Sigma)$ and $\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)'$, the exact distribution of Hotelling's T^2 statistic $T^2 = (\bar{X}_n - \mu)'(\frac{1}{n}\hat{\Sigma})^{-1}(\bar{X}_n - \mu)$ is known as *Hotelling's T^2 distribution with parameters k and $n-1$*. This distribution is none other than a rescaled F-distribution; $\frac{n-k}{(n-1)k}T^2$ follows the F-distribution with k and $n-k$ degrees of freedom. Again, this converges

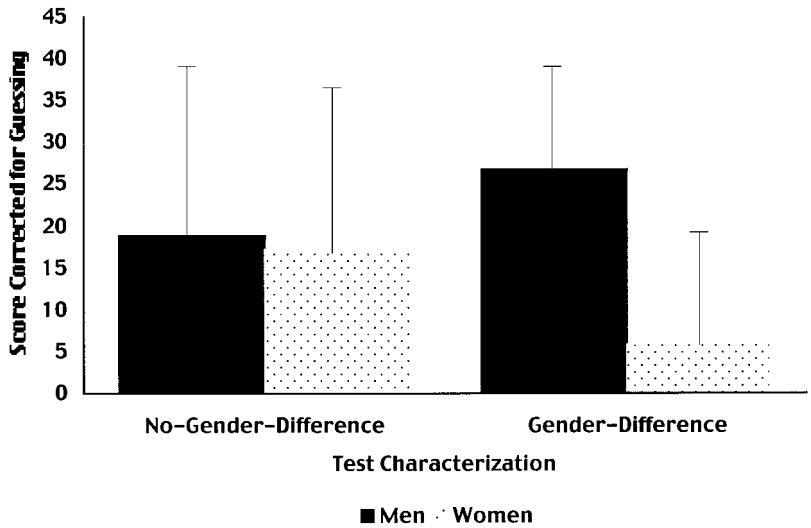


FIGURE 5.6. Average math test scores for men and women are not different in the group that was told that the test produces no difference; average score for women is significantly lower in the group that was told that the test produces gender differences. [**SSQ99**, Fig. 2].

to a χ^2 distribution with k degrees of freedom as $n \to \infty$ but the smaller the sample size, the more conservative the inference should be.

In social science, such a small sample situation arises, e.g., when the experiment is small in psychology or when each observation is at a state level in economics. At any rate, it cannot be stressed more that these distributions are valid only when each X_i follows an i.i.d. normal distribution.[8] If you are not comfortable with the normality assumption and still have a small sample size, then replacing the critical value of a normal distribution (or of a χ^2 distribution) by that of a t-distribution (or of a T^2-distribution) does not justify your inference procedure. In that case, you may have to make some nonnormal distributional assumption, or you may need to adopt some correction methods such as the Edgeworth expansion or the Berry–Esseen theorem (Proposition 3.6).

EXAMPLE 5.1 (Stereotype threat). In social psychology, stereotype threat refers to a situation in which an individual faces judgment based on the stereotypes of the group to which they belong. [**SSQ99**] examine the stereotype threat that female college students face when they perform math. The authors selected 30 women and 24 men at the University of Michigan to take a math test. The subjects were randomly divided into two groups: (1) the first group was told before the exam that the test had been shown to produce gender differences and (2) the second group was told that the test had been shown to produce no gender difference. If stereotype threat exists, female students in group 1 would perform worse than in group 2. The test score X_i of an individual i is assumed to follow $N(\mu + \gamma \mathbb{1}\{i \text{ is female}\} + \delta \mathbb{1}\{i \text{ is in group } 1\} +$

[8]For more detailed relationship of the t-statistic and the t-distribution, see [**YFK07**].

$\theta \mathbb{1}\{i \text{ is female and in group } 1\}, \sigma^2)$. The parameter θ is of interest and can be estimated by the difference of the average scores of female students in groups 1 and 2 while σ^2 by the sample variance of all students. Then, the t-statistic $\frac{\hat{\theta}-\theta}{\hat{\sigma}/\sqrt{n}}$ follows a t-distribution with $n-4$ degrees of freedom (this is a special case of Proposition 7.9). The authors then test a hypothesis $H_0 : \theta = 0$ and find that $\hat{\theta}$ is significantly negative with a p-value of 0.0462.[9] In fact, women's average math score was as high as men's in group 2 ($\gamma \approx 0$) and men's average score was not very different in two groups ($\delta \approx 0$), whereas women's score in group 1 was significantly lower (Figure 5.6). The authors conclude that "these findings provide strong evidence that women's underperformance on these difficult math tests results from stereotype threat, rather than from sex-linked ability differences that are detectable only on advanced math material." The paper sparked massive literature on stereotype threat, which now provides mixed conclusions; for a recent meta-analysis, see [**FW15**].

[9]They compare t^2 to an F-distribution instead of t to a t-distribution, but these are equivalent when a point hypothesis is concerned.

CHAPTER 6

Maximum Likelihood Estimation

All models are wrong, but some are useful.

George E. P. Box, 1978

6.1. The Principle of Maximum Likelihood

The *principle of maximum likelihood* is the principle of ordinariness. Suppose we throw a coin 100 times and get 82 heads. There are two ways to interpret this outcome. First, we can think that we are extremely lucky, since getting at least 82 heads in 100 coin tosses can happen only about twice in a billion trials. Or instead, we can choose to think that the coin we're tossing may be heavily biased, since then what just happened to us is anything but surprising. In particular, if the probability of the coin landing on heads is 82%, the given outcome becomes the most ordinary. This latter interpretation is the philosophy behind maximum likelihood estimation (MLE).

But why does that work? Why is regarding ourselves as ordinary better than regarding ourselves as special? To see this, it is important to distinguish the ordinariness of one observation from that of the whole sample. Note that each individual observation can be as ordinary or as rare as it can be. However, if rare events happen as rarely as they are and ordinary events happen as ordinarily as they are, then that whole dataset—collectively—is merely ordinary. In other words, if 10% of the outcomes are as rare as 10%, 20% of them as rare as 20%, 30% as 30%, and so on, then that is the most "ordinary" outcome we can expect as one big chunk of observations, which is exactly what Proposition 4.2 predicts to happen eventually, that the frequency counts converge to the true probability distribution as the sample size increases. Therefore, seeing the entire dataset as the most *collectively* ordinary outcome makes more and more sense as we include more and more data points. Ergo, we are exploiting the fundamental characteristics of the probability that, in the long run, only the most ordinary event can take place.

Thus, in MLE, we seek the value of the parameter that renders the whole observations the least surprising. This is a generic principle that can be applied widely from parametric to nonparametric models, but its mathematics is the easiest to illustrate for parametric models, which we focus on here. In economics and business research, a parametric model of an economic agent's decision may be implied by economic theory, and the parameters thereof may be estimated by MLE.

We inherit the i.i.d. setup from Section 4.3. In particular, let $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ be a parametric model with the density p_θ on a common support, the score $\dot{\ell}_\theta$ continuous in both θ and x, and the invertible Fisher information matrix I_θ. The likelihood of the sample is given by $\prod_{i=1}^n p_\theta(X_i)$. Maximizing it is equivalent to maximizing its logarithm, $\ell_{n,\theta}(X) = \sum_{i=1}^n \ell_\theta(X_i)$. Thus, the maximum likelihood estimator (MLE) solves[1]

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^n \ell_\theta(X_i).$$

This method is mathematically justified by the following property.

Theorem 6.1 (Identification)**.** *If $\theta \neq \theta_0$ implies $p_\theta \neq p_{\theta_0}$, then $\mathbb{E}_{\theta_0}[\ell_\theta(X_i)]$ is uniquely maximized at $\theta = \theta_0$.*

Proof. A slightly stronger claim is proved in [**vdV98**, Lemma 5.35]. ■

MLE is an example of M-estimation (Example 4.3). However, since the log likelihood is assumed to be differentiable here, it can be reduced to a Z-estimation problem (Example 4.2)

$$\dot{\ell}_{n,\hat{\theta}}(X) = \sum_{i=1}^n \dot{\ell}_{\hat{\theta}}(X_i) = 0,$$

which is justified in light of $\mathbb{E}_{\theta_0}[\dot{\ell}_{\theta_0}(X_i)] = 0$ (Theorem 4.6).

6.2. As the Construction Method for Efficient Estimators

MLE is strongly motivated in relation to the Cramér–Rao bound. Recall that the crucial step in Theorem 4.7 is the following Cauchy–Schwarz inequality (simplified for a univariate case)

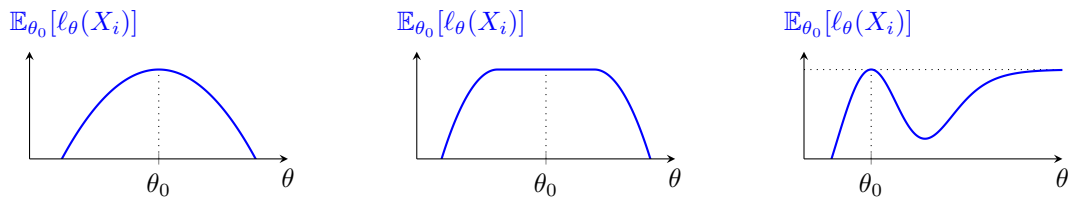
$$\operatorname{Cov}(\hat{\theta}, \dot{\ell}_{n,\theta})^2 \leq \operatorname{Var}(\hat{\theta}) \operatorname{Var}(\dot{\ell}_{n,\theta}) \qquad \Longleftrightarrow \qquad |\operatorname{Corr}(\hat{\theta}, \dot{\ell}_{n,\theta})| \leq 1.$$

Ergo, the attainability of efficiency hinges on the attainability of this as an equality. In other words, $\hat{\theta}$ is efficient if and only if $\hat{\theta}$ and $\dot{\ell}_\theta$ are linearly dependent, that is, $\dot{\ell}_{n,\theta}(X) = a_\theta(\hat{\theta}(X) - b_\theta)$ for some a_θ and b_θ. If $\hat{\theta}$ is unbiased, then b_θ must be θ for the score to have mean zero, hence $\dot{\ell}_{n,\theta}(X) = a_\theta(\hat{\theta}(X) - \theta)$.

Note that if we substitute θ by $\hat{\theta}$, we get $\dot{\ell}_{n,\hat{\theta}} = a_{\hat{\theta}}(\hat{\theta} - \hat{\theta}) = 0$, which coincides with the Z-estimation formulation of MLE. This implies that the efficient estimator of a smooth parametric model is none other than the MLE itself. However, the converse does not hold—there are cases where the MLE exists but no efficient estimator does. This is because the MLE is neither necessarily unbiased nor necessarily linearly dependent on the score. For that matter, the MLE can exist even when the MVU estimator does not.

The bottom line is that, if we want an efficient estimator, try MLE and check if it is unbiased and its variance attains the bound. If so, bingo. But even if not, we then know that there does not exist an efficient estimator.

[1] The acronym "MLE" stands for both "maximum likelihood estimation" and "maximum likelihood estimator," depending on the context.



(A) Likelihood that has an isolated maximum.

(B) Likelihood that has a plateaued maximum.

(C) Likelihood that asymptotes to its maximum.

FIGURE 6.1. The identification condition in Theorem 6.2 requires that $\sup_{\|\theta-\theta_0\|\geq\varepsilon} \mathbb{E}_{\theta_0}[\ell_\theta(X_i)] < \mathbb{E}_{\theta_0}[\ell_{\theta_0}(X_i)]$ for every $\varepsilon > 0$. It is satisfied in (A) but not in (B) or (C).

EXAMPLE 6.1 (Normal). Suppose $X_i \sim N(\mu, \sigma^2)$ where μ is known. Then

$$\dot{\ell}_{n,\sigma^2}(X) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^4} = \frac{n}{2\sigma^4}\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 - \sigma^2\right].$$

Therefore, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$ attains the Cramér–Rao bound.[2] If μ is not known, the MLE for σ^2 is equal to the plug-in, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\bar{X}_n)^2$, which neither is unbiased nor attains the Cramér–Rao bound. However, a MLE exists in either case.

EXAMPLE 6.2 (Cauchy). Suppose X_i follows the Cauchy distribution with pdf $p_\theta(x) = \pi^{-1}[1+(x-\theta)^2]^{-1}$. Then, the MLE for θ exists and is unbiased but does not attain the Cramér–Rao bound.

Example 6.1 provokes a thought that, since μ is consistently estimable, $\hat{\sigma}^2$ "approaches" an efficient estimator as we have more and more samples. This intuition is correct—the MLE of a smooth parametric model is *asymptotically* efficient, be it efficient in finite samples or not. The bottom line is, the converse *does* hold asymptotically for smooth parametric models.

6.3. Asymptotic Efficiency and Inference

The Cramér–Rao bound is a finite-sample result and is not attainable in many examples. However, the intuition of the Cramér–Rao bound goes beyond finite samples. In fact, even when MLE is not efficient in finite samples, it is asymptotically efficient (Definition 4.7) for all smooth models. This section presents the formal results for this.

First, MLE is consistent when it is identified and satisfies some regularity conditions. The identification condition requires that the function have an isolated maximum (Figure 6.1). The second condition states that the sample log likelihood must converge to the population log likelihood uniformly over the parameter space.

Theorem 6.2 (Consistency of MLE). *Let $\theta_0 \in \Theta$ be the true parameter. Suppose that $\sup_{\|\theta-\theta_0\|\geq\varepsilon} \mathbb{E}_{\theta_0}[\ell_\theta(X_i)] < \mathbb{E}_{\theta_0}[\ell_{\theta_0}(X_i)]$ for every $\varepsilon > 0$ and $\sup_{\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^n \ell_\theta(X_i) -$*

[2]Note that since μ is known, $\hat{\sigma}^2$ is unbiased without any correction (Example 4.4).

$\mathbb{E}_{\theta_0}[\ell_\theta(X_i)]| \to^p 0$. *Then, the maximum likelihood estimator* $\hat\theta$ *converges in probability to* θ_0.

PROOF. It follows from [vdV98, Theorem 5.7]. The argument is as follows.

$$\begin{aligned}\mathbb{E}_{\theta_0}[\ell_{\theta_0}(X_i)] - \mathbb{E}_{\theta_0}[\ell_{\hat\theta}(X_i)] &= \mathbb{E}_{\theta_0}[\ell_{\theta_0}(X_i)] - \tfrac{1}{n}\sum \ell_{\theta_0}(X_i) + \tfrac{1}{n}\sum \ell_{\theta_0}(X_i) - \mathbb{E}_{\theta_0}[\ell_{\hat\theta}(X_i)] \\ &\le \mathbb{E}_{\theta_0}[\ell_{\theta_0}(X_i)] - \tfrac{1}{n}\sum \ell_{\theta_0}(X_i) + \tfrac{1}{n}\sum \ell_{\hat\theta}(X_i) - \mathbb{E}_{\theta_0}[\ell_{\hat\theta}(X_i)] \\ &\le 2\sup_{\theta\in\Theta}\left|\tfrac{1}{n}\sum \ell_\theta(X_i) - \mathbb{E}_{\theta_0}[\ell_\theta(X_i)]\right| \xrightarrow{p} 0,\end{aligned}$$

where the first inequality uses the maximizing property of $\hat\theta$ and the last convergence uses the second assumption. By the first assumption, it follows that $\hat\theta \to^p \theta_0$. ■

The following result shows that MLE is asymptotically efficient, so its variance converges to the inverse of the Fisher information matrix.

Theorem 6.3 (Parametric efficiency of MLE)**.** *Suppose* θ_0 *is in the interior of* Θ *and there exists a measurable function* m *such that* $\mathbb{E}_{\theta_0}[m(X_i)^2] < \infty$ *and for every* θ_1 *and* θ_2 *in a neighborhood of* θ_0*,*

$$|\ell_{\theta_1}(x) - \ell_{\theta_2}(x)| \le m(x)\|\theta_1 - \theta_2\|.$$

Suppose also that P_θ *is smooth as described in Section 4.3 and* $\ell_{n,\theta}$ *is twice differentiable with the second derivative continuous in both* θ *and* x*. If* $\hat\theta$ *is consistent,*

$$\sqrt{n}(\hat\theta - \theta_0) = I_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot\ell_{\theta_0}(X_i) + o_P(1) \rightsquigarrow N(0, I_{\theta_0}^{-1}).$$

PROOF. I give a sketch. A rigorous proof is found in [vdV98, Theorem 5.39]. By Taylor's theorem,[3]

$$0 = \frac{1}{\sqrt{n}}\sum \dot\ell_{\hat\theta}(X_i) = \frac{1}{\sqrt{n}}\sum \dot\ell_{\theta_0}(X_i) + \frac{1}{\sqrt{n}}\sum \ddot\ell_{\theta_0}(X_i)(\hat\theta - \theta_0) + o_P(1).$$

This rearranges as

$$\sqrt{n}(\hat\theta - \theta_0) = -\left(\frac{1}{n}\sum \ddot\ell_{\theta_0}\right)^{-1}\frac{1}{\sqrt{n}}\sum \dot\ell_{\theta_0}(X_i) + o_P(1).$$

Observe that $\frac{1}{n}\sum \ddot\ell_{\theta_0} \to^p -I_{\theta_0}$ by Theorem 4.6 and $\frac{1}{\sqrt{n}}\sum \dot\ell_{\theta_0} \rightsquigarrow N(0, I_{\theta_0})$ by the CLT. Then the result follows by Slutsky's lemma. ■

The "practical" version of the statement is $\hat\theta \rightsquigarrow N(\theta, \frac{1}{n}I_{\theta_0}^{-1})$. When MLE is efficient in finite samples, its mean is θ and variance is $I_{n,\theta_0}^{-1} = \frac{1}{n}I_{\theta_0}^{-1}$ (Definition 4.4). Theorem 6.3 then adds that the shape of the distribution approaches normal. Moreover, even when MLE is not efficient (so no efficient estimator exists), the mean and variance of MLE converge to a hypothetical normally distributed efficient estimator.

[3]The hard part is to show that the remainder is indeed $o_P(1)$.

Remark 6.1. The assumption of everywhere twice continuous differentiability of the likelihood can be relaxed to accommodate some kinky and jumpy distributions, which may appear in censored models, auctions, or corporate finance [**vdV98**, Section 5.5].

EXAMPLE 6.3 (Exponential distribution). Let $X = (X_1, \dots, X_n)$ be i.i.d. observations from an exponential distribution with pdf $p_\lambda(x) = \lambda e^{-\lambda x} \mathbb{1}\{x > 0\}$. Then we have $\dot{\ell}_\lambda(x) = \lambda^{-1} - x$ and $\ddot{\ell}_\lambda(x) = -\lambda^{-2}$, so $I_\lambda = \lambda^{-2}$. Since the log likelihood of the sample is $\ell_{n,\lambda}(X) = n \log \lambda - \lambda \sum_{i=1}^n X_i$, the maximum likelihood estimator is $\hat{\lambda} = \bar{X}_n^{-1}$. Now, it is easy to see that this estimator is not unbiased, much less efficient. However, by Theorem 6.3 we know that it is *asymptotically* efficient, i.e., $\sqrt{n}(\hat{\lambda} - \lambda) \rightsquigarrow N(0, \lambda^2)$.

EXERCISE 6.1 (Logistic location model). Let $X = (X_1, \dots, X_n)$ be i.i.d. observations from a logistic distribution with pdf $p_\mu(x) = \frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2}$, where we assume $s > 0$ is known. Find the asymptotic distribution of MLE for μ.

EXERCISE 6.2 (Superconsistency). The counterexample to the Cramér–Rao bound (Example 4.12) also works as a counterexample for MLE. We have i.i.d. $X_i \sim U[0, \theta]$, so $\hat{\theta} = X_{(n)}$. Show that $\hat{\theta}$ converges faster than $\sqrt{n}$.

EXERCISE 6.3 (Boundary). Let $X = (X_1, \dots, X_n)$ be i.i.d. $N(\mu, 1)$ with $\mu \geq 0$. Show that the maximum likelihood estimator is $\hat{\mu} = \bar{X}_n \vee 0$ and that $\sqrt{n}(\hat{\mu} - \mu_0)$ does not converge to $N(0, I_{\mu_0}^{-1})$ when $\mu_0 = 0$. Which assumption in Theorem 6.3 does this violate?

EXERCISE 6.4 (Incidental parameter problem). MLE may not work when there is a growing number of parameters, a situation that arises quite often in panel data models. For illustration, consider $X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} \sim N(\left[\begin{smallmatrix} \mu_i \\ \mu_i \end{smallmatrix}\right], \sigma^2 \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right])$, where the parameter μ_i defines the means of X_{i1} and X_{i2} but is not related to any other observation; such parameters that only concern a finite number of observations despite the growing n are called *incidental*. Derive MLEs for μ_i and σ^2 and verify that they are not even consistent. Which assumption in Theorem 6.2 does this violate?

To carry out inference with a maximum likelihood estimator, we need its asymptotic variance, which is the inverse Fisher information matrix. For a normal location model with known variance, we can analytically calculate that $I_\mu^{-1} = \Sigma$. When the Fisher information depends on unknown quantities, we will need to estimate it in some way. In Example 6.3, the asymptotic variance is given as λ^2, which depends on unknown λ. Since $\hat{\lambda}$ is consistent, the plug-in estimator $\hat{\lambda}^2$ converges in probability to λ^2 by the continuous mapping theorem. In general, a more simplistic plug-in method works just fine.

Theorem 6.4 (Fisher information estimation)**.** *Suppose there exists a measurable function m such that $\mathbb{E}_{\theta_0}[m(X_i)] < \infty$ and for every θ_1 and θ_2 in a neighborhood of θ_0,*

$$\|\dot{\ell}_{\theta_1}(x)\dot{\ell}_{\theta_1}(x)' - \dot{\ell}_{\theta_2}(x)\dot{\ell}_{\theta_2}(x)'\| \leq m(x)\|\theta_1 - \theta_2\|.$$

If $\hat{\theta}$ is consistent, then we have $\frac{1}{n}\sum_{i=1}^n \dot{\ell}_{\hat{\theta}}(X_i)\dot{\ell}_{\hat{\theta}}(X_i)' \to^p I_{\theta_0}$.

PROOF. As in [vdV98, Example 19.7], the L_1-bracketing number of (each component of) $\{\dot{\ell}_\theta \dot{\ell}_\theta' : \|\theta - \theta_0\| < \varepsilon\}$ is finite for sufficiently small $\varepsilon > 0$. Since $\mathbb{E}_{\theta_0}[m(X_i)] < \infty$, this class of functions is P_{θ_0}-Glivenko–Cantelli by [vdV98, Theorem 19.4]. Then the claim follows by the consistency of $\hat{\theta}$. ■

EXAMPLE 6.4 (Huff model). Large stores tend to be located farther away from residential areas, and small stores closer. The Huff model describes how consumers choose which store to go to. Suppose that there are S stores in a neighborhood of interest. The Huff model postulates that, building on some axioms, consumer i chooses to shop at store s with probability

$$\frac{z(s)/d(s)^\lambda}{\sum_{\tilde{s}=1}^S z(\tilde{s})/d(\tilde{s})^\lambda},$$

where $z(s)$ is the size of store s, $d(s)$ the distance to store s, and λ the parameter that measures the consumer's preference for distance over store size. If $\lambda = 0$, the consumer chooses solely on the basis of the size of the store; if $\lambda = \infty$, they always go to the closest store. We can also interpret λ as an effective geographic market size; the larger λ is, the smaller is the geographic market region. If λ is finite, they visit each store with positive probability, so the smoothness condition in Section 4.3 including the common support is met.[4]

Consider surveying consumers in a given neighborhood (so that $d(s)$ is the same across all consumers), and label the stores they visited by $X_1, \dots, X_n$. Then, the log likelihood for each shopping trip X_i is given by

$$\ell_\lambda(X_i) = \log \frac{z(X_i)/d(X_i)^\lambda}{\sum_{s=1}^S z(s)/d(s)^\lambda}.$$

The score and Hessian are

$$\dot{\ell}_\lambda(X_i) = -\log d(X_i) + \frac{\sum_{s=1}^S z(s) d(s)^{-\lambda} \log d(s)}{\sum_{s=1}^S z(s) d(s)^{-\lambda}},$$

$$\ddot{\ell}_\lambda(X_i) = -\frac{\sum_{s=1}^S z(s) d(s)^{-\lambda} (\log d(s))^2}{\sum_{s=1}^S z(s) d(s)^{-\lambda}} + \left[\frac{\sum_{s=1}^S z(s) d(s)^{-\lambda} \log d(s)}{\sum_{s=1}^S z(s) d(s)^{-\lambda}}\right]^2.$$

Since the Hessian does not depend on the data, we obtain the Fisher information as $I_\lambda = -\ddot{\ell}_\lambda$ in light of Theorem 4.6, which can then be estimated by $I_{\hat{\lambda}}$. In practice, the value of λ may vary depending on the purpose of shopping, e.g., for regular shopping and for fill-in shopping, as well as on the neighborhood; [HSA72] estimate the model separately for the purpose of shopping and neighborhood and find that there is not much variation of λ across different neighborhoods of middle class while there is a significant variation of λ across different purposes of shopping: the major stores, the second stores, and the fill-in stores (Table 6.1).

Remark 6.2. Sometimes, we have a part of observations that does not depend on θ. Suppose that an observation consists of a pair (X, Y) and their joint distribution is given by $p_{X,Y;\theta}(x, y) = p_X(x) p_\theta(y \mid x)$. This is to say that the parameter θ affects

[4] If it helps, you can think that z and s are continuous functions on $\mathbb{R}$.

TABLE 6.1. Preference for distance in the Huff model [HSA72, p. 157]

	Neighborhood											
	Model Cities			3rd Ward			19th Ward			Maplewood		
Stores	$\hat{\lambda}$	[80% CI]		$\hat{\lambda}$	[80% CI]		$\hat{\lambda}$	[80% CI]		$\hat{\lambda}$	[80% CI]	
Major	1.06	0.72	1.39	0.17	0.02	0.36	0.93	0.46	1.41	0.49	0.00	0.97
Second	1.10	0.24	4.30	1.17	0.54	1.82	0.64	0.00	1.27	0.58	0.27	1.20
Fill-in	1.71	1.25	2.21	1.75	1.29	2.23	1.61	1.04	2.25	0.15	0.01	1.00

the conditional distribution of Y given X but not the marginal distribution of X. Such X is called the covariates. In this case, the log likelihood can be split into the marginal log likelihood of X and the conditional log likelihood of Y given X, i.e., $\ell_\theta(x, y) = \log p_X(x) + \log p_\theta(y \mid x)$. Note that the first term is independent of θ and hence does not influence the maximization. This means that MLE can be carried out *without specifying the marginal distribution of* X. Moreover, since the derivative of $\log p_X(x)$ with respect to θ is zero, we can ignore the marginal of X in the score and Fisher information calculation as well. This is an obvious extension of MLE to a particular type of semiparametric models that have covariates. An example of this is the logistic regression (Chapter 8).

6.4. Misspecification and Quasi-Maximum Likelihood

What if our guess about the parametric model is incorrect? Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sample from a probability distribution P_0 such that $P_0 \neq P_\theta$ for every $\theta \in \Theta$. In social science, it is reasonable to expect that no model is fully correct to its fine details, so this view matches many researchers' perception. In this case, MLE is specifically called *quasi-maximum likelihood estimation (QMLE)* and estimates a pseudo-parameter that maximizes the misspecified log likelihood, i.e.,

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E}_{P_0}[\ell_\theta(X_i)],$$

which equivalently minimizes the Kullback–Leibler divergence from P_{θ_0} to P_0. If the model is sufficiently smooth and identified, we expect that the score still satisfies the moment equality

$$\mathbb{E}_{P_0}[\dot{\ell}_{\theta_0}(X_i)] = 0$$

as the first-order condition for the maximization, but now under P_0 instead of P_{θ_0}. Meanwhile, the variance of the score $\mathbb{E}_{P_0}[\dot{\ell}_{\theta_0}(X_i)\dot{\ell}_{\theta_0}(X_i)']$ is different from either the Fisher information I_{θ_0} or the expected second derivative $-\mathbb{E}_{P_0}[\ddot{\ell}_{\theta_0}(X_i)]$ since the proof of Theorem 4.6 does not go through with $\mathbb{E}_{P_0}[\ddot{p}_\theta/p_\theta] \neq \int \ddot{p}_\theta$.

The consequence of this is that, while the maximum likelihood estimator can still be consistent (to the pseudo-parameter thusly defined) and asymptotically normal, its asymptotic variance will no longer be equal to the inverse variance of the score.

Below are the conditions for consistency and asymptotic normality of QMLE.

Theorem 6.5 (Consistency of QMLE)**.** *Let $\theta_0 \in \Theta$ be the pseudo-parameter that maximizes $\mathbb{E}_{P_0}[\ell_\theta(X_i)]$. Suppose that $\sup_{\|\theta-\theta_0\|\geq\varepsilon} \mathbb{E}_{P_0}[\ell_\theta(X_i)] < \mathbb{E}_{P_0}[\ell_{\theta_0}(X_i)]$ for every $\varepsilon > 0$ and $\sup_{\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^n \ell_\theta(X_i) - \mathbb{E}_{P_0}[\ell_\theta(X_i)]| \to^p 0$. Then, the quasi-maximum likelihood estimator $\hat{\theta}$ converges in probability to θ_0.*

Proof. It follows from [vdV98, Theorem 5.7]. ■

Theorem 6.6 (Asymptotic distribution of QMLE)**.** *Suppose θ_0 is in the interior of Θ and there exists a measurable function m such that $\mathbb{E}_{P_0}[m(X_i)^2] < \infty$ and for every θ_1 and θ_2 in a neighborhood of θ_0,*

$$\|\ell_{\theta_1} - \ell_{\theta_2}(x)\| \leq m(x)\|\theta_1 - \theta_2\|.$$

Suppose also that P_θ is smooth as described in Section 4.3 and $\mathbb{E}_{P_0}[\ddot{\ell}_{\theta_0}(X_i)]$ is invertible. If $\hat{\theta}$ is consistent, then

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) &= -\mathbb{E}_{P_0}[\ddot{\ell}_{\theta_0}(X_i)]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + o_P(1)\\ &\rightsquigarrow N\Big(0,\, \mathbb{E}_{P_0}[\ddot{\ell}_{\theta_0}(X_i)]^{-1}\mathbb{E}_{P_0}[\dot{\ell}_{\theta_0}(X_i)\dot{\ell}_{\theta_0}(X_i)']\mathbb{E}_{P_0}[\ddot{\ell}_{\theta_0}(X_i)]^{-1}\Big).\end{aligned}$$

Proof. It follows from [vdV98, Theorem 5.23]. ■

Note that this asymptotic variance reduces to $I_{\theta_0}^{-1}$ if the model is correctly specified and the likelihood is twice differentiable. Therefore, when conducting inference with MLE, it is advisable to always use an estimator of this variance formula instead of the simplified inverse Fisher information, since then the statistical inference is valid regardless of the presence of misspecification. The following theorem gives one such estimator.

Theorem 6.7 (Variance estimation for QMLE)**.** *Suppose that there exist measurable functions m_1 and m_2 such that $\mathbb{E}_{P_0}[m_1(X_i)] < \infty$, $\mathbb{E}_{P_0}[m_2(X_i)] < \infty$, and for every θ_1 and θ_2 in a neighborhood of θ_0,*

$$\begin{aligned}\|\dot{\ell}_{\theta_1}(x)\dot{\ell}_{\theta_1}(x)' - \dot{\ell}_{\theta_2}(x)\dot{\ell}_{\theta_2}(x)'\| &\leq m_1(x)\|\theta_1 - \theta_2\|,\\ \|\ddot{\ell}_{\theta_1}(x) - \ddot{\ell}_{\theta_2}(x)\| &\leq m_2(x)\|\theta_1 - \theta_2\|.\end{aligned}$$

If $\hat{\theta}$ is consistent, then we have $\frac{1}{n}\sum_{i=1}^n \dot{\ell}_{\hat{\theta}}(X_i)\dot{\ell}_{\hat{\theta}}(X_i)' \to^p \mathbb{E}_{P_0}[\dot{\ell}_{\theta_0}(X_i)\dot{\ell}_{\theta_0}(X_i)']$ and $\frac{1}{n}\sum_{i=1}^n \ddot{\ell}_{\hat{\theta}}(X_i) \to^p \mathbb{E}_{P_0}[\ddot{\ell}_{\theta_0}(X_i)]$.

Proof. Essentially the same as Theorem 6.4. ■

There are some cases where the pseudo-parameter withstands our test of interpretability. Suppose we are interested in the mean of X_i and postulate that X_i is i.i.d. normal. We may be wrong in our choice of normality, but the mean of X_i is still well defined and can be of interest. In this sense, $\theta_0 = \mathbb{E}_{P_0}[X_i]$ upholds the interpretation as the "true" parameter. Since the maximum likelihood estimator of the normal location model is the sample average $\hat{\theta} = \bar{X}_n$, it consistently estimates the true mean anyway.

This observation generalizes to more complicated models, specifically to ones where the conditional mean is of interest and the deviation therefrom is modeled as normal. A prominent example in economics is *limited information maximum likelihood (LIML)*, the details of which are left for more advanced courses.

EXERCISE 6.5 (Logistic location model). Let $X = (X_1, \ldots, X_n)$ be as defined in Exercise 6.1. However, suppose that we use a normal location model with unit variance, that is, we mistakingly model it as $N(\mu, 1)$. Verify that the misspecification-robust variance in Theorem 6.6 equals the variance of X_i. Also, compute how much efficiency loss we incur for this misspecification relative to Exercise 6.1.

EXERCISE 6.6 (Logistic scale model). Let $X = (X_1, \ldots, X_n)$ be i.i.d. standard logistic random variables. Suppose that we misspecify the model as a normal scale model, $N(0, \sigma^2)$ with parameter σ^2. Derive the pseudo-parameter σ_0^2 to which the QMLE asymptotes.

EXERCISE 6.7 (Lognormal). Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sample of the chi-square distribution with four degrees of freedom (Example 2.8). However, suppose we postulate a lognormal model with an unknown location parameter μ and a known scale parameter σ (Example 2.9). Derive the pseudo-parameter μ to which the QMLE asymptotes, and derive the asymptotic distribution of the QMLE. You may use the following fact: If $X \sim \chi^2(4)$, $\mathbb{E}[\log X] = 1 + \log 2 - \gamma$ and $\mathrm{Var}(\log X) = \frac{\pi^2}{6} - 1$ where $\gamma \approx 0.577$ is Euler's constant.

6.5. Wald, Likelihood Ratio, and Lagrange Multiplier Tests

Now that we have the maximum likelihood estimator $\hat{\theta}$ that is asymptotically normal and an estimator of its asymptotic variance that is consistent, we may carry out statistical inference as we learned in Chapter 5. On top of that, there are two more ways to carry out inference in the MLE framework that are worthy of discussion.

The simple application of Chapter 5 leads to the Wald test, which measures the extremeness of our sample relative to the hypothesis by the distance of $\hat{\theta}$ to θ_0. Alternatively, we can measure the extremeness by the difference of the log likelihood at $\hat{\theta}$ and θ_0, leads to what is known as the likelihood ratio test. Moreover, we may take the slope of the log likelihood at θ_0 as the extremeness measure, since the slope of the population log likelihood must be zero at θ_0 if the hypothesis is true; this leads to the Lagrange multiplier test. These tests are illustrated in Figure 6.2.

Under correct specification, the three tests become asymptotically equivalent. Nevertheless, it is valuable to understand them all since each one of them can be generalized to different models other than MLE. In this section, we assume correct specification for simplicity, but it is straightforward to extend the Wald and Lagrange multiplier tests to misspecification [**Whi82**].

The *Wald test* is based on an estimator $\hat{\theta}$ and relies on the asymptotic normality $\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N(0, I_{\theta_0}^{-1})$ in Theorem 6.3 to construct a valid test.

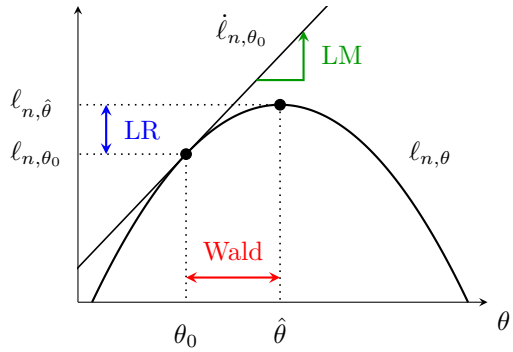


FIGURE 6.2. Three tests. The Wald test examines if $\hat{\theta}$ is close to the null θ_0. The LR test examines if the log-likelihood at the null is close to the maximized log-likelihood. The LM test examines if the slope at the null is close to 0.

Theorem 6.8 (Wald test)**.** *Suppose $\hat{I}_{\theta_0}$ is a consistent estimator for I_{θ_0}. Under the assumptions of Theorem 6.3, we have*

$$W := n(\hat{\theta} - \theta_0)'\hat{I}_{\theta_0}(\hat{\theta} - \theta_0) = n(\hat{\theta} - \theta_0)'I_{\theta_0}(\hat{\theta} - \theta_0) + o_P(1) \rightsquigarrow \chi^2(k).$$

PROOF. Write $W = \sqrt{n}(\hat{\theta} - \theta_0)'I_{\theta_0}\sqrt{n}(\hat{\theta} - \theta_0) + \sqrt{n}(\hat{\theta} - \theta_0)'(\hat{I}_{\theta_0} - I_{\theta_0})\sqrt{n}(\hat{\theta} - \theta_0)$. The first term converges to $\chi^2(k)$ by the CMT (Theorem 3.3) and the second is $O_P(1)o_P(1)O_P(1) = o_P(1)$. ■

The *likelihood ratio (LR) test* is based on the M-estimation formulation of MLE and checks if the attained maximum value of the objective function $\sum \ell_{\hat{\theta}}(X_i)$ is close to the value of the objective function at the null $\sum \ell_{\theta_0}(X_i)$. Their difference can be shown to converge in distribution to a chi-square distribution, whence we can construct a valid test.

Theorem 6.9 (Likelihood ratio test)**.** *Under the assumptions of Theorem 6.3 and additional smoothness, we have*

$$LR := 2\left(\sum_{i=1}^n \ell_{\hat{\theta}}(X_i) - \sum_{i=1}^n \ell_{\theta_0}(X_i)\right) = n(\hat{\theta} - \theta_0)'I_{\theta_0}(\hat{\theta} - \theta_0) + o_P(1) \rightsquigarrow \chi^2(k).$$

PROOF. As in Theorem 6.3, I only give intuition. A rigorous proof requires modified versions of [**vdV98**, Theorem 7.2 and Lemma 19.31].

By Taylor's theorem,

$$LR = -2\sum \dot{\ell}_{\hat{\theta}}(X_i)'(\theta_0 - \hat{\theta}) - [\sqrt{n}(\theta_0 - \hat{\theta})]'\tfrac{1}{n}\sum \ddot{\ell}_{\hat{\theta}}(X_i)[\sqrt{n}(\theta_0 - \hat{\theta})] + o_P(1).$$

The first term is zero since $\hat{\theta}$ solves $\sum \dot{\ell}_{\hat{\theta}}(X_i) = 0$. The second term converges to $\chi^2(k)$ by the CMT. ■

The *Lagrange multiplier (LM) test* or the *score test* is based on the M-estimation formulation of MLE and uses if the score at the null $\sum \dot{\ell}_{\theta_0}(X_i)$ is close to 0. Since

this converges in distribution to a normal by the CLT, we can construct a valid test therefrom.

Theorem 6.10 (Lagrange multiplier test)**.** *Suppose $\hat{I}_{\theta_0}$ is a consistent estimator for I_{θ_0}. Then, under the assumptions of Theorem 6.3,*

$$LM := \left(\sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i)\right)' (n\hat{I}_{\theta_0})^{-1} \left(\sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i)\right) \rightsquigarrow \chi^2(k).$$

Under the assumptions of Theorem 6.3, we have $LM = n(\hat{\theta} - \theta_0)' I_{\theta_0}(\hat{\theta} - \theta_0) + o_P(1)$.

Proof. Write $LM = (\frac{1}{\sqrt{n}} \sum \dot{\ell}_{\theta_0})' I_{\theta_0}^{-1} (\frac{1}{\sqrt{n}} \sum \dot{\ell}_{\theta_0}) + (\frac{1}{\sqrt{n}} \sum \dot{\ell}_{\theta_0})' (\hat{I}_{\theta_0}^{-1} - I_{\theta_0}^{-1}) (\frac{1}{\sqrt{n}} \sum \dot{\ell}_{\theta_0})$. The first term converges to $\chi^2(k)$ by the CLT (Theorem 3.5) and CMT and the second is $o_P(1)$. To obtain the second representation, note that $\sum \dot{\ell}_{\theta_0} = \sum \dot{\ell}_{\theta_0} - \sum \dot{\ell}_{\hat{\theta}} = \sum \ddot{\ell}_{\theta_0}(X_i)(\hat{\theta} - \theta_0) + o_P(1)$ as in Theorem 6.3. ■

The name comes from the fact that the test statistic is given as the Lagrange multiplier of the M-estimation problem constrained by the null hypothesis

$$\max_{\theta} \ \ell_{n,\theta}(X) \quad \text{subject to} \quad \theta = \theta_0.$$

Although this problem is degenerate in that θ is fully specified by the constraint, the test generalizes to the case where the hypothesis specifies part of θ or the range of θ. In general, if the constraint is not binding (that is, the constrained maximum equals the unconstrained maximum), the Lagrange multiplier is zero.

Asymptotically, the log likelihood approaches a quadratic function whose curvature matches the inverse variance of $\hat{\theta}$, and the three tests become all equivalent.

Corollary 6.11 (Trinity)**.** *Let t_1 and t_2 be any two of the three test statistics W, LR, and LM. Under the assumptions of their respective validity, we have $t_1 - t_2 \to^p 0$.*

In sum, the Wald test examines if the alternative exhibits the property of the null, so there is no need to evaluate anything for the null; the LM test examines if the null exhibits the property of the true model, and there is no need to carry out estimation of θ (if the null hypothesis specifies the entirety of θ); the LR test evaluates both hypotheses and compares how close they are. The Wald test can be extended to arbitrary models that yield an estimator and is often the default choice in many cases. The LM test extends to general Z-estimation problems and is a preferred choice when estimation is computationally burdensome. The LR test is extendable to some M-estimation problems such as optimally-weighted GMM [**Hay00**, Section 7.4] and can be applied to the test of overidentification [**Hay00**, Section 8.5].

In the context of linear regression, it is known that the following relationship holds in finite samples [**Bre79**],

$$W \geq LR \geq LM.$$

This means that Wald is the least conservative and LM is the most conservative. Unlike what we see in Section 7.2.4, however, it is not so customary to use—for the sake of finite-sample conservatism—the LM test when the Wald test is available.

CHAPTER 7

Linear Regression

Of all the principles . . . there is no more general, more exact, and more easy of application [than that] . . . which consists of rendering the sum of squares of the errors a minimum.

Adrien-Marie Legendre, translated by H. A. Ruger and H. M. Walker, 1805

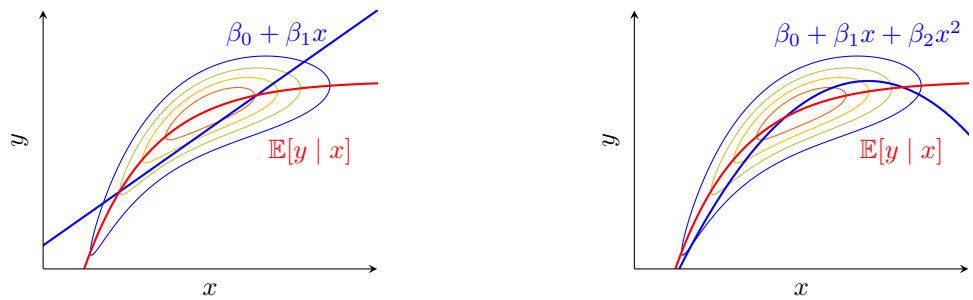
Linear regression plays a fundamental role in quantitative research in social science. Historical origins aside, the adjective "linear" should be understood as "linear in parameters" but not "linear in variables." A caveat on notation. We have so far denoted random variables by capital letters (e.g., X) and nonrandom values by small letters (e.g., x). When we discuss linear regression in econometrics, it is customary to use small letters for each observation of random variables (e.g., x_i) and capital letters for the vector or matrix of the entire sample (e.g., $X = (x_1, \ldots, x_n)'$). We hereafter conform to this convention.

7.1. Introduction

7.1.1. Regression vs. classification. Both regression and classification refer to statistical methods of explaining a random variable by other random variables. The variable to be explained is called the *dependent variable*, *outcome variable*, or *response variable*, and the variables to explain it are called the *regressors*, *explanatory variables*, *independent variables*, etc. Distinction between regression and classification lies in the type of the dependent variable. When the dependent variable is numerical (continuous), we call it *regression*; when categorical (discrete), *classification*. This nomenclature is mostly agreed upon in computer science, while statisticians use "regression" as an umbrella term that encompasses both. For example, the canonical classification method is sadly called logistic regression (Chapter 8).

7.1.2. Two projections. When predicting y with the knowledge of x, the best predictor in squared loss is the conditional expectation

$$\mathbb{E}[y \mid x] = \arg\min_g \mathbb{E}[(y - g(x))^2],$$



(A) $\mathbb{E}[y \mid x]$ and the linear projection of y onto x.

(B) $\mathbb{E}[y \mid x]$ and the linear projection of y onto x and x^2.

FIGURE 7.1. Two projections. The contour map shows the joint pdf of (x, y). The red line is the conditional expectation of y given x. The blue line is the linear projection of y onto functions of x.

where g runs through all measurable functions (Theorem 2.11). We can think of this as the "measure-theoretic" projection of y onto x.

When we restrict the class of predictors to linear functions of x, the best predictor is the linear projection of y onto x, i.e.,

$$\mathbb{E}[xx']^{-1}\mathbb{E}[xy] = \arg\min_b \mathbb{E}[(y - x'b)^2].$$

This is the "analytic" projection of y onto x in the Hilbert space $L_2(P)$.

Since the first projection is minimizing over a wider class of functions, its minimized loss is smaller than or equal to the second projection's (Figure 7.1A). On the other hand, we can augment the vector x by polynomials, $(x, x_1^2, \dots, x_1x_2, \dots)$, and approximate any analytic function g by a linear function thereof (Figure 7.1B).[1] Therefore, distinction between the two is not as clear-cut as it might first seem. Moreover, when the two coincide—when we get the basis right—nice properties hold both statistically and interpretationally.

The derivative of the conditional expectation $\partial\mathbb{E}[y \mid x]/\partial x$ is called the *partial effect* of x on y and measures how much y changes in expectation in response to a unit change in x, *holding other variables constant* [**Gre18**, Chapter 3]. Such ceteris paribus consideration is key in causal inference, and when the two projections coincide, β holds the interpretation as the partial effect of x on y; for example, β_1 is the partial effect of x_1 holding all other variables constant.

For every $b \in \mathbb{R}^k$, we can without loss of generality write

$$y = x'b + \varepsilon_b$$

for $\varepsilon_b := y - x'b$. The second projection finds the value of b that makes the second moment of ε_b smallest. This choice of b and ε_b is specifically denoted by β and ε, i.e.,

$$\beta := \mathbb{E}[xx']^{-1}\mathbb{E}[xy], \qquad \varepsilon := y - x'\beta.$$

[1] We can also choose other bases. For example, trigonometric series span all square-integrable functions on a bounded domain.

Being the residual of an orthogonal projection, ε has the nice characterization by

$$\mathbb{E}[x\varepsilon] = 0,$$

where 0 is understood as a $k \times 1$ vector of zeros. This means that if x contains a constant element known as the *intercept*, we also have $\mathbb{E}[\varepsilon] = 0$.

Suppose we have n observations of (y_i, x_i), each of which satisfies

$$y_i = x_i'\beta + \varepsilon_i, \qquad \mathbb{E}[x_i\varepsilon_i] = 0.$$

We denote this by $Y = X\beta + \mathcal{E}$ where

$$\underset{(n\times 1)}{Y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \underset{(n\times k)}{X} := \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}, \qquad \underset{(n\times 1)}{\mathcal{E}} := \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

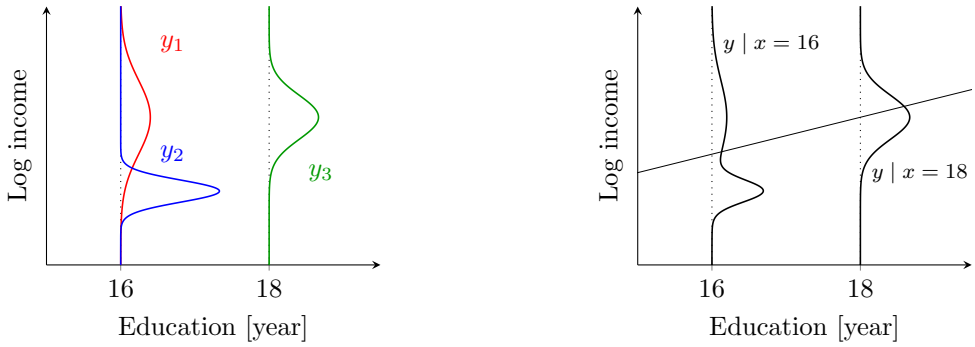
Note that each *row* of X is the transpose of the *column* vector x_i.

7.1.3. Which projection do we have in mind in practice? When prediction is our concern, the first projection is of interest. Since x is known at the time we make a prediction about y, we want to exploit the information from x as much as possible, be it linear or not. So the conditional expectation is what we are after.

When causal inference is our concern, we still want the first projection. Since x is to be intervened in, we do know x and want to know how much we should adjust x in order to attain a desired change in y.

Thus, what is practically of interest is in many cases the conditional expectation, not the linear projection. However, as estimating the conditional expectation is often hard, many practitioners see the linear projection as a convenient interpretation as an approximation to the conditional expectation function. In other words, we may not believe in the strong assumptions that make $x'\beta$ the conditional expectation of y given x, but even in their violation, $x'\beta$ can be interpreted as the best linear approximation to the conditional expectation. In particular, by the law of iterated expectations (Theorem 2.9), we can write $\beta = \mathbb{E}[xx']^{-1}\mathbb{E}[x\mathbb{E}[y \mid x]]$. This tells that β is a linear projection of the conditional expectation $\mathbb{E}[y \mid x]$ onto x. Such a property—when strong assumptions hold, we have strong interpretation; when they fail but weak assumptions hold, we have weak interpretation—is often preferred by economists. However, this perception should not be taken without reservations. For example, when we intervene in x, the distribution of x changes, which in turn changes this projection (but not the measure-theoretic projection). So how should we decide on the degree of intervention from β? If we intervene in a specific range of x, what does β say about the causal effect?

Another appeal of the linear projection is its compatibility with asymptotic theory. We will see in the next section that weak assumptions are enough to yield statements about the asymptotic properties, while strong assumptions are required for similar results in finite-sample terms.



(A) The i.n.i.d. view of the data. (B) The i.i.d. view of the data.

FIGURE 7.2. Plausibility of the i.i.d. sampling assumption.

7.1.4. On the sampling assumption. In this chapter, we make the assumption of i.i.d. sampling. While this assumption can be relaxed, it is also practically important to see how far we can go with this assumption.

Consider investigating the relationship between the education level x measured by the number of years in school and the productivity level y measured by the logarithm of the income. We have three observations $i = 1, 2, 3$. The first two individuals are college graduates, so $x_1 = x_2 = 16$, and the third one is an MBA holder, i.e., $x_3 = 18$. There are diverse occupations taken by college graduates, so the possible range of their income is extensive. Say, the individual $i = 1$ took a bachelor in math and works for a hedge fund, and her income level is high but also its variation is substantial as half of her income is a bonus tied to the performance of their fund (the red distribution in Figure 7.2A). The individual $i = 2$ majored literature and works as a middle school teacher, and her income is relatively lower than $i = 1$ but with much lower volatility (the blue distribution). The last individual $i = 3$ is a consultant, and her income is high with a little less volatility than $i = 1$ (the green distribution). In this view, the distribution of (x, y) is different for each individual, though they are not correlated. In statistics terminology, the observations are independent but not identically distributed (i.n.i.d.).

On the other hand, if we don't have the data on their majors or occupations, we have no way to distinguish these distributions. From an economist's point of view, the data look like Figure 7.2B; the income distribution for $x = 16$ is dispersed, and that for $x = 18$ is higher on average and more concentrated. Indeed, if we regard that we randomly sample individuals from the pool of all individuals, their (x, y) is not only independent but "identically distributed" as the aggregate distribution in Figure 7.2B. Thus, the assumption of i.i.d. sampling can be maintained even when the distribution of each observation seems to vary much.

With this interpretation, we can draw the projection line for predicting the income from education (the black line in Figure 7.2B). This line is meaningful in that when we want to predict a new individual with a college degree but without information of

her major or occupation (just as in the original dataset), then this line gives our best prediction for her income; the same thing for a new individual with an MBA.

Note that, while the joint distribution of (y_i, x_i) can be considered i.i.d. across i, the conditional distribution of y_i given x_i cannot be deemed i.i.d. across x_i without losing much generality. It is only sensible to expect that the income distribution takes a very different shape for high school graduates, college graduates, and graduate degree holders, beyond the difference in the mean. This distributional variation is known as heteroskedasticity.

But what if we observe data on their occupations? If so, we can distinguish (part of) the distributions in Figure 7.2A, which can be incorporated by the means of more explanatory variables. We can still maintain the i.i.d. sampling assumption as long as we allow the distribution of the income to now depend on *both* the education level and occupation.

Finally, note that the key component in this "reinterpretation" is the fact that the observations are independent. If the observations exhibit interrelation such as time series, spatial, or network dependence, it is generally hard or impossible to recover the i.i.d. sampling interpretation.

Remark 7.1. Microsoft Word thinks that "-skedasticity" should be spelled as "-scedasticity." In economics, there has been a consensus that it should be spelled with a k long before Word was created [**McC85**]. Some say that this is a difference between American English and British English. It is only ironic that *Biometrika* uses a c and *Econometrica* (mostly) uses a k [**DE01**, Appendix B].

7.2. Theory of Ordinary Least Squares

The *ordinary least squares (OLS)* minimizes the sample counterpart of the mean squared error,

$$\sum_{i=1}^{n}(y_i - x_i'b)^2 = (Y - Xb)'(Y - Xb)$$

with respect to $b \in \mathbb{R}^k$. The FOC gives $-2X'(Y - Xb) = 0$, which solves

$$\hat{\beta} = (X'X)^{-1}(X'Y).$$

We denote by $\hat{\mathcal{E}} := Y - X\hat{\beta}$ the residuals, and by $\hat{Y} := X\hat{\beta}$ the fitted values.

Remark 7.2. The unobservable remainder ε_i is called the *error term* while the estimated remainder $\hat{\varepsilon}_i$ is called the *residual*. For arbitrary x, the value $x'\hat{\beta}$ is called the *predicted value* of y; it is specifically called the *fitted value* of y if x is taken from observed values.

There are broadly two types of assumptions that give different properties, the *unconditional restrictions* and *conditional restrictions*. The unconditional restrictions are weaker than the conditional, and are enough to give nice asymptotic properties. The conditional restrictions are strong enough to give nice finite-sample properties.

7.2.1. Asymptotic properties. In a nutshell, we have only one essential assumption for OLS to work (Assumption 7.1). Moreover, the i.i.d. assumption can be relaxed, e.g., to accommodate time series dependence.

Assumption 7.1 (Essential assumption)**.** $\{y_i, x_i\}$ are i.i.d. with finite second moments. $\mathbb{E}[x_i x_i']$ is invertible and $\mathbb{E}[\varepsilon_i^2 x_i x_i']$ exists.

The next assumption is not at all necessary and in fact is too strong for economic applications, but if imposed, it can simplify the asymptotic variance formula a little bit. Failure to satisfy Assumption 7.2 is called *unconditional heteroskedasticity*.

Assumption 7.2 (Unconditional homoskedasticity)**.** $\mathbb{E}[\varepsilon_i^2 x_i x_i'] = \mathbb{E}[\varepsilon_i^2]\mathbb{E}[x_i x_i']$.

With these assumptions, we can infer consistency and asymptotic normality of OLS, whose proof is also worth knowing.

Theorem 7.1 (Asymptotic normality)**.** *Under Assumption 7.1, $\hat{\beta}$ is consistent and*

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow N\Big(0,\, \mathbb{E}[x_i x_i']^{-1}\mathbb{E}[\varepsilon_i^2 x_i x_i']\mathbb{E}[x_i x_i']^{-1}\Big).$$

Under Assumption 7.2, the asymptotic variance formula reduces to $\mathbb{E}[\varepsilon_i^2]\mathbb{E}[x_i x_i']^{-1}$.

Proof. Since $\mathbb{E}[xx']$ is invertible, $X'X$ is invertible with probability approaching 1. Using $\hat{\beta} = (X'X)^{-1}[X'(X\beta + \mathcal{E})] = \beta + (X'X)^{-1}(X'\mathcal{E})$, we can write

$$\sqrt{n}(\hat{\beta} - \beta) = (\tfrac{1}{n}X'X)^{-1}\tfrac{1}{\sqrt{n}}X'\mathcal{E}.$$

By the LLN (Theorem 3.4) and the CMT (Theorem 3.3), $(\frac{1}{n}X'X)^{-1} \to^p \mathbb{E}[xx']^{-1}$. By the CLT (Theorem 3.5), $\frac{1}{\sqrt{n}}X'\mathcal{E} \rightsquigarrow N(0, \mathbb{E}[\varepsilon^2 xx'])$. Then asymptotic normality (and consistency) follows by Slutsky's lemma (Theorem 3.2). The last claim is trivial. ■

Moreover, OLS can even be the best possible estimator for estimating the population linear projection line.

Proposition 7.2 (Semiparametric efficiency)**.** *Under Assumption 7.1, $\hat{\beta}$ is semiparametrically efficient.*

Proof. Apply a similar argument as [**vdV98**, Example 25.28]. ■

Remark 7.3. Beware that the notion of semiparametric efficiency depends on the scope of probability distributions we deem possible. Therefore, if we *add* conditions such as Assumption 7.3, Proposition 7.2 may no longer hold.[2] See Section 7.3.

7.2.2. Finite-sample properties. Similar results in finite-sample (nonasymptotic) terms can be derived with stronger assumptions. As the regression analysis started with small-sample analysis, there is rich theory on finite-sample properties, and it is no surprise that many textbooks start with and devote a fair amount of space to the derivation of finite-sample results. They are of less interest to economists, however, who enjoy large samples and dislike strong assumptions.

[2]By the way, adding Assumption 7.2 does not impair the validity of Proposition 7.2.

Assumption 7.3 (Conditional linearity)**.** $\mathbb{E}[y_i \mid X] = x_i'\beta$, or equivalently, $\mathbb{E}[\varepsilon_i \mid X] = 0$. $X'X$ is invertible.

Assumption 7.3 requires that the two projections coincide. This assumption not only yields finite-sample properties but also admits improvement of the asymptotic efficiency bound (Section 7.3). Note that $y_i = x_i'\beta + \varepsilon_i$ is always well defined as a linear projection, so there was no need to assume "unconditional linearity" in Section 7.2.1. That is, there was no counterpart of Assumption 7.3 for Theorem 7.1 (except that invertibility of $\mathbb{E}[x_i x_i']$ is a counterpart of invertibility of $X'X$).

Assumption 7.4 (Conditional homoskedasticity)**.** $\mathbb{E}[\varepsilon_i^2 \mid X] = \sigma^2$ for some constant σ^2 and $\mathbb{E}[\varepsilon_i \varepsilon_j \mid X] = 0$ for $i \neq j$.

Failure to satisfy Assumption 7.4 is called *conditional heteroskedasticity*. Assumptions 7.3 and 7.4 follow from the following stronger condition.

Assumption 7.5 (Conditional normality)**.** $\mathcal{E} \mid X \sim N(0, \sigma^2 I)$ for some scalar σ^2 and an $n \times n$ identity matrix I.

Exercise 7.1. Show that Assumption 7.4 implies Assumption 7.2 and that Assumption 7.5 implies Assumptions 7.3 and 7.4.

Example 7.1 (Unconditionally homoskedastic but not conditionally so). Let x and z be independent Rademacher random variables, that is, 1 with probability 1/2 and -1 with probability 1/2. Let $\varepsilon = z\mathbb{1}\{x = 1\}$. Then $\mathbb{E}[\varepsilon \mid x] = 0$. Since $x^2 \equiv 1$, we have $\mathbb{E}[\varepsilon^2 x^2] = \mathbb{E}[\varepsilon^2]\mathbb{E}[x^2]$. However, $\mathbb{E}[\varepsilon^2 \mid x] = \mathbb{1}\{x = 1\}$ depends on x.

Remark 7.4. $\mathbb{E}[\varepsilon_i^2 \mid X] = \sigma^2$ for some constant σ^2 implies $\sigma^2 = \mathbb{E}[\varepsilon_i^2]$, but $\mathbb{E}[\varepsilon_i^2 x_i x_i'] = \sigma^2 \mathbb{E}[x_i x_i']$ for some constant σ^2 does *not* imply $\sigma^2 = \mathbb{E}[\varepsilon_i^2]$.

Remark 7.5. Under Assumption 7.3, $\mathbb{E}[\varepsilon_i^2 \mid X] = \mathrm{Var}(\varepsilon_i \mid X)$, so $\mathbb{E}[\varepsilon_i^2 \mid X] = \text{constant}$ if and only if $\mathrm{Var}(\varepsilon_i \mid X) = \text{constant}$.

The following result is a finite-sample counterpart of Theorem 7.1. Under stronger assumptions, OLS is unbiased and normally distributed in finite samples.

Proposition 7.3 (Finite-sample unbiasedness and normality)**.** *Under Assumption 7.3,* $\mathbb{E}[\hat{\beta} \mid X] = \beta$. *Under Assumptions 7.3 and 7.4,* $\mathrm{Var}(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}$ *and* $\mathrm{Cov}(\hat{\beta}, \hat{\mathcal{E}} \mid X) = 0$. *Under Assumption 7.5,* $\hat{\beta} \mid X \sim N(\beta, \sigma^2 (X'X)^{-1})$.

Proof. Unbiasedness follows from $\mathbb{E}[\hat{\beta} \mid X] - \beta = (X'X)^{-1} X' \mathbb{E}[\mathcal{E} \mid X] = 0$. Next, $\mathrm{Var}(\hat{\beta} \mid X) = \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \mid X] = (X'X)^{-1} X' \mathbb{E}[\mathcal{E}\mathcal{E}' \mid X] X (X'X)^{-1} = \sigma^2 (X'X)^{-1}$. For the third claim, note that $Y - X\hat{\beta} = [I - X(X'X)^{-1}X']Y = [I - X(X'X)^{-1}X']\mathcal{E}$ where I stands for an identity matrix of a conformable size. Then, $\mathrm{Cov}(\hat{\beta}, \hat{\mathcal{E}} \mid X) = \mathbb{E}[(\hat{\beta} - \beta)(Y - X\hat{\beta})' \mid X] = \mathbb{E}[(X'X)^{-1} X' \mathcal{E}\mathcal{E}' [I - X(X'X)^{-1}X'] \mid X] = \sigma^2 \mathbb{E}[(X'X)^{-1} X' [I - X(X'X)^{-1}X'] \mid X] = 0$. The last claim is trivial given $\hat{\beta} = \beta + (X'X)^{-1} X' \mathcal{E}$. ■

Under these strong assumptions, optimality of OLS also extends to finite samples. The following is a finite-sample counterpart of Proposition 7.2.

Proposition 7.4 (Gauss–Markov)**.** *Under Assumptions 7.3 and 7.4, $\hat{\beta}$ is the best linear unbiased estimator (BLUE), that is,* $\mathrm{Var}(\hat{\beta} \mid X) \leq \mathrm{Var}(\tilde{\beta} \mid X)$ *for every unbiased estimator $\tilde{\beta}$ that is linear in Y.*

Proof. Let $\tilde{\beta} = C(X)Y$ and define $D = C(X) - (X'X)^{-1}X'$. Write $\tilde{\beta} = \hat{\beta} + DY = \hat{\beta} + DX\beta + D\mathcal{E}$. Since $\mathbb{E}[D\mathcal{E} \mid X] = 0$ and $\tilde{\beta}$ is unbiased, we have $DX = 0$. Therefore, $\tilde{\beta} = \hat{\beta} + D\mathcal{E}$, and hence $\mathrm{Var}(\tilde{\beta} \mid X) = \mathrm{Var}(\hat{\beta} \mid X) + \mathrm{Cov}(\hat{\beta}, D\mathcal{E} \mid X) + \mathrm{Cov}(D\mathcal{E}, \hat{\beta} \mid X) + \mathrm{Var}(D\mathcal{E} \mid X)$. Now, $\mathrm{Cov}(D\mathcal{E}, \hat{\beta} \mid X) = \mathbb{E}[D\mathcal{E}\mathcal{E}'X(X'X)^{-1} \mid X] = \mathbb{E}[\varepsilon_i^2]DX(X'X)^{-1} = 0$ since $DX = 0$. Then, the result follows since $\mathrm{Var}(\hat{\beta} \mid X) - \mathrm{Var}(\tilde{\beta} \mid X) = \mathrm{Var}(D\mathcal{E} \mid X)$ is positive semidefinite. ■

Exercise 7.2. Show that $\mathrm{Var}(\hat{\beta}) \leq \mathrm{Var}(\tilde{\beta})$ under Proposition 7.4. *Hint: Use Theorem 2.13.*

Remark 7.6. The OLS estimator can be viewed as a (conditional) MLE under Assumptions 7.1 and 7.5. The simplified variance formula $\mathbb{E}[\varepsilon_i^2]\mathbb{E}[x_i x_i']^{-1}$ in Theorem 7.1 corresponds to the correctly specified variance formula in Theorem 6.3, and the heteroskedasticity-robust variance $\mathbb{E}[x_i x_i']^{-1}\mathbb{E}[\varepsilon_i^2 x_i x_i']\mathbb{E}[x_i x_i']^{-1}$ to the misspecification-robust variance in Theorem 6.6. This correspondence is useful, e.g., when we interpret regularization as a Bayesian inference.

Remark 7.7. [**Han22**] shows that the requirement of linearity can be lifted, that is, that OLS is the best unbiased estimator (BUE), among a certain class of estimators. See also [**LW23**] for the latest discussion of the results.

7.2.3. Standard errors. Heteroskedasticity does not affect the estimation of β, that is, the formula $\hat{\beta} = (X'X)^{-1}(X'Y)$ still stands. Moreover, the asymptotic variance formula $\mathbb{E}[xx']^{-1}\mathbb{E}[\varepsilon^2 xx']\mathbb{E}[xx']^{-1}$ is valid either in the homoskedastic or heteroskedastic case. It is just that homoskedasticity admits a minor simplification. Therefore, as long as we estimate the asymptotic variance using the general formula, it is consistent to the correct variance regardless of the presence of heteroskedasticity. This estimator is also known as the Huber–White standard error.

Theorem 7.5 (Heteroskedasticity-robust standard error)**.** *Under Assumption 7.1 and $\mathbb{E}[\|x_i\|^4] < \infty$,*

$$\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i'\right)\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \xrightarrow{p} \mathbb{E}[x_i x_i']^{-1}\mathbb{E}[\varepsilon_i^2 x_i x_i']\mathbb{E}[x_i x_i']^{-1}.$$

Proof. By the LLN (Theorem 3.4), $\frac{1}{n}\sum xx' \to^p \mathbb{E}[xx']$. With $\varepsilon - \hat{\varepsilon} = x'(\hat{\beta} - \beta)$,

$$\hat{\varepsilon}^2 xx' = [\varepsilon - x'(\hat{\beta} - \beta)]^2 xx' = \varepsilon^2 xx' - 2[x'(\hat{\beta} - \beta)]\varepsilon xx' + [x'(\hat{\beta} - \beta)]^2 xx'.$$

First, $\frac{1}{n}\sum \varepsilon^2 xx' \to^p \mathbb{E}[\varepsilon^2 xx']$ by the LLN. Second, using $|x'(\hat{\beta} - \beta)| \leq \|x\|\|\hat{\beta} - \beta\|$,

$$\|\tfrac{1}{n}\textstyle\sum[x'(\hat{\beta} - \beta)]\varepsilon xx'\| \lesssim \|\hat{\beta} - \beta\| \cdot \tfrac{1}{n}\textstyle\sum|\varepsilon|\|x\|^3 = O_P(\tfrac{1}{\sqrt{n}}) \cdot O_P(1) = o_P(1),$$

since $\mathbb{E}[|\varepsilon|\|x\|^3] \le \sqrt{\mathbb{E}[\varepsilon^2\|x\|^2]\mathbb{E}[\|x\|^4]} < \infty$ by the Cauchy–Schwarz inequality. Third,

$$\|\tfrac{1}{n}\textstyle\sum[x'(\hat\beta-\beta)]^2 xx'\| \lesssim \|\hat\beta-\beta\|^2 \cdot \tfrac{1}{n}\textstyle\sum\|x\|^4 = O_P(\tfrac{1}{n})\cdot O_P(1) = o_P(1).$$

Then, the result follows by the CMT (Theorem 3.3) and Slutsky's lemma (Theorem 3.2). ■

Despite this general formula, most statistical software reports as default the following estimator that is valid only under homoskedasticity. One reason may be Proposition 7.9. Some software such as Microsoft Excel does not even have an option to report the robust one.

Proposition 7.6 (Default standard error)**.** *Under Assumption 7.1,*[3]

$$\left(\frac{1}{n}\sum_{i=1}^n \hat\varepsilon_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \xrightarrow{p} \mathbb{E}[\varepsilon_i^2]\mathbb{E}[x_i x_i']^{-1}.$$

Exercise 7.3. Prove Proposition 7.6.

We can think of a regression version of Bessel's correction for the conditional variance. The following suggests using the divisor $n-k$ in place of n, where k is the dimension of β. This is referred to as *small-sample correction* or *finite-sample adjustment*.

Proposition 7.7 (Small-sample correction)**.** *Under Assumptions 7.3 and 7.4,*

$$\mathbb{E}\left[\frac{1}{n-k}\sum_{i=1}^n \hat\varepsilon_i^2 \,\middle|\, X\right] = \sigma^2,$$

$$\sigma^2 A \le \mathbb{E}\left[\frac{1}{n-k}\sum_{i=1}^n \hat\varepsilon_i^2 x_{ij}x_{ir} \,\middle|\, X\right] \le \sigma^2 B,$$

where A is the average of $n-k$ smallest elements of $\{x_{ij}x_{ir}\}_{i=1}^n$ and B of $n-k$ largest elements.

Proof. Observe that $\hat{\mathcal{E}} = M_X\mathcal{E}$ where $M_X := I - X(X'X)^{-1}X'$ is the projection matrix onto the orthocomplement of the space spanned by X. Thus, $\mathbb{E}[\hat{\mathcal{E}}'\hat{\mathcal{E}} \mid X] = \mathbb{E}[\operatorname{tr}(\mathcal{E}'M_X\mathcal{E}) \mid X] = \operatorname{tr}(\mathbb{E}[\mathcal{E}\mathcal{E}' \mid X]M_X) = \mathbb{E}[\varepsilon^2]\operatorname{tr}(M_X)$. Since the trace of a projection matrix is the dimension of the target space and X has rank k, we have $\operatorname{tr}(M_X) = n-k$. Therefore, $\frac{1}{n-k}\mathbb{E}[\hat{\mathcal{E}}'\hat{\mathcal{E}} \mid X] = \mathbb{E}[\varepsilon^2]$.

Let X_j be the jth column of X and diag the operator that converts a vector into a diagonal matrix. Then, $\mathbb{E}[\sum \hat\varepsilon_i^2 x_{ij}x_{ir} \mid X] = \mathbb{E}[\operatorname{tr}(\mathcal{E}'M_X \operatorname{diag}(X_j)\operatorname{diag}(X_r)M_X\mathcal{E}) \mid X] = \mathbb{E}[\varepsilon^2]\operatorname{tr}(M_X\operatorname{diag}(X_j)\operatorname{diag}(X_r))$. We can write $M_X = U\Lambda U'$ for U orthogonal and Λ diagonal with eigenvalues of M_X. Then, the trace is bounded by the sums of $n-k$ largest and smallest diagonal elements of $\operatorname{diag}(X_j)\operatorname{diag}(X_r)$ by [**CR09**, Theorem 4.1 and Corollary 4.2]. ■

There are a few more ways to correct for the possible bias in the variance estimator. See [**AP09**, Chapter 8] for further reading.

[3]For this to be a valid asymptotic variance of $\hat\beta$, we need Assumption 7.2.

Remark 7.8. It is often observed in practice that the robust standard errors are larger than the default ones. However, this relationship is not a theorem. Observe that $\mathbb{E}[xx']^{-1}\mathbb{E}[\varepsilon^2 xx']\mathbb{E}[xx']^{-1} - \mathbb{E}[\varepsilon^2]\mathbb{E}[xx']^{-1} = \mathbb{E}[xx']^{-1}\operatorname{Cov}(\varepsilon^2, xx')\mathbb{E}[xx']^{-1}$. So, the default standard errors can be larger if $\operatorname{Cov}(\varepsilon^2, xx') < 0$.

7.2.4. Inference. With an estimator of the asymptotic variance, we can conduct statistical inference on β. Usually, we are interested in whether the coefficient on one regressor is nonzero; for example, when we want to examine the effect of schooling on wages, we want to test whether the coefficient on schooling is nonzero. Also, we are often interested in whether the entire set of regressors has any power of explaining the dependent variable. In this case we want to test if *all* coefficients are zero. Such hypotheses are given by a hypothesis of the form $H_0 : \beta_j = 0$ or $H_0 : \beta = 0$. More generally, we develop a method to test

$$H_0 : \beta_j = b_j \qquad \text{and} \qquad H_0 : C\beta = b,$$

where C is an $(r \times k)$ matrix of full row rank for some $r \le k$ and b an $(r \times 1)$ vector. Obviously, the former is a special case of the latter, but we state the results for the two cases separately. Let

$$\hat{V} := \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\frac{1}{n-k}\sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i'\right)\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1},$$
$$\hat{V}_0 := \frac{1}{n}\left(\frac{1}{n-k}\sum_{i=1}^n \hat{\varepsilon}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}.$$

Correction by $n-k$ is unnecessary for the asymptotic validity of testing since $(n-k)/n \to 1$ for fixed k, but it is customary to use it nonetheless. In Stata, the `reg` command uses $\hat{V}_0$ by default (e.g., `reg y x`), and the option `vce(r)` switches to $\hat{V}$ (e.g., `reg y x, vce(r)`).

Theorem 7.8 (Asymptotic testing)**.** *For a full row-rank matrix C with rank r, let*

$$t_j := \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{jj}}}, \quad F_C := \frac{(C\hat{\beta} - C\beta)'(C\hat{V}C')^{-1}(C\hat{\beta} - C\beta)}{r}.$$

Under Assumption 7.1 and $\mathbb{E}[\|x_i\|^4] < \infty$, $t_j \rightsquigarrow N(0,1)$ and $rF_C \rightsquigarrow \chi^2(r)$.

Proof. It follows from Theorems 7.1 and 7.5. ■

Remark 7.9. Without $\mathbb{E}[\|x_i\|^4] < \infty$ but with Assumption 7.2, we may replace $\hat{V}$ by $\hat{V}_0$.

The statistic t_j is called the *t-statistic* for the hypothesis $H_0 : \beta_j = b_j$ and F_C the *F-statistic* for the hypothesis $H_0 : C\beta = b$. For example, the hypothesis $H_0 : (\beta_0, \beta_1) = (1, 2)$ is translated as

$$C = \begin{bmatrix} 1 & & 0 & \cdots & 0 \\ & 1 & 0 & \cdots & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

It can also test a hypothesis on a linear combination, e.g., $H_0 : 2\beta_0 - \beta_1 = 3$ is identified as

$$C = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \end{bmatrix}, \qquad b = 3.$$

In most statistical software, the t-statistics and the F-statistic for the hypotheses $H_0 : \beta_j = 0$ and $H_0 : \beta_{-0} = 0$ are given as part of the standard output, where β_{-0} stands for the vector of coefficients *excluding the intercept if exists*. This latter hypothesis can be interpreted as the hypothesis that all regressors are meaningless in predicting y_i. To see this, observe that under this hypothesis, the regression equation reduces to $y_i = \beta_0 + \varepsilon_i$ if the intercept is included and to $y_i = \varepsilon_i$ if not. In either case, it is saying that predicting y_i using the information of x_i is equivalent to predicting y_i without using any information. Thus, under this hypothesis, the regressors collectively have no predictive power on y_i.

Example 7.2 (Hedge funds). Hedge funds are seeking to produce returns that are not affected by the performance of the whole stock market, which they call the *absolute return*. Let y_i be the excess return of a hedge fund and x_i be the excess return of S&P 500, which is a proxy for the market return. To test whether the hedge fund return is orthogonal to the market return, you run the regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Which hypothesis do you want to test? What are the values of C and b?

Exercise 7.4 (Index funds). The index fund is a mutual fund that aims to reproduce the benchmark return such as S&P 500 and Nasdaq. Let y_i be the excess return of an S&P 500 index fund and x_i be the excess return of S&P 500. To assess the quality of the index fund, you run the regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Which hypothesis do you want to test? What are C and b?

Exercise 7.5 (Buffett's alpha). Let y_i be the excess return of Warren Buffett's portfolio. To know the source of his alpha, you decide to decompose his returns into the factors proposed in the Fama–French three-factor model: the market factor measured by the market excess return (MER), the size factor measured by the small-minus-big market capitalization (SMB), and the value factor measured by the high-minus-low book-to-market ratio (HML). You run the regression $y_i = \beta_0 + \beta_1 \mathrm{MER}_i + \beta_2 \mathrm{SMB}_i + \beta_3 \mathrm{HML}_i + \varepsilon_i$. To investigate if Buffett's return can be explained by any of these factors, which hypothesis do you test? What are C and b? What about the hypothesis that there is Buffett's alpha not explicable by these factors?

Remark 7.10. The statistic rF_C is called the *Wald statistic*. The chi-square test can be easily extended to nonlinear smooth hypotheses [**Hay00**, Proposition 2.3].

A practice that is as popular as using the small-sample correction of the standard errors is to use the critical values from a t- or F-distribution in lieu of a normal or chi-square distribution. Its justification relies on the strong assumption of conditional linearity and normality that hardly any practitioner believes.[4] However, the critical values thusly constructed are slightly more conservative than—and converge in the limit to—the ones from a normal or chi-square and hence give practitioners a (false)

[4]Note also that the variance used to normalize the test statistics must be the default one.

sense of security that the conclusions are not entirely based upon the asymptotic approximation.[5]

Proposition 7.9 (Finite-sample testing)**.** *For a full row-rank matrix C with rank r, let*

$$t_{0,j} := \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{0,jj}}}, \quad F_{0,C} := \frac{(C\hat{\beta} - C\beta)'(C\hat{V}_0 C')^{-1}(C\hat{\beta} - C\beta)}{r}.$$

Under Assumptions 7.3 and 7.5, $t_{0,j} \mid X \sim t(n-k)$ and $F_{0,C} \mid X \sim F(r, n-k)$, where $t(m)$ is the t-distribution with m degrees of freedom and $F(r,m)$ the F-distribution with r and m degrees of freedom.

Proof. Write

$$t_{0,j} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 (X'X)^{-1}_{jj}}} \cdot \left(\frac{1}{n-k} \frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{\sigma^2} \right)^{-1/2}.$$

The first fraction follows $N(0,1)$ conditional on X. Since $\frac{\mathcal{E}}{\sigma} \mid X \sim N(0, I)$ and M_X is symmetric and idempotent with rank $n-k$, we have that $\frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{\sigma^2} = \frac{\mathcal{E}}{\sigma}' M_X \frac{\mathcal{E}}{\sigma}$ follows $\chi^2(n-k)$ conditional on X. Next, being linear functions of $\mathcal{E}$, both $\hat{\beta}$ and $\hat{\mathcal{E}}$ are jointly normal conditional on X. Since $\mathrm{Cov}(\hat{\beta}, \hat{\mathcal{E}} \mid X) = 0$ by Proposition 7.3, we have $\hat{\beta} \perp \hat{\mathcal{E}} \mid X$. In a nutshell, $t_{0,j} = A/\sqrt{\frac{B}{n-k}}$ where $A \mid X \sim N(0,1)$, $B \mid X \sim \chi^2(n-k)$, and $A \perp B \mid X$. By the definition of the t-distribution, $t_{0,j} \mid X \sim t(m)$.

Next, write

$$F_{0,C} = \frac{(C\hat{\beta} - C\beta)'[\sigma^2 C(X'X)^{-1}C']^{-1}(C\hat{\beta} - C\beta)}{r} \cdot \left(\frac{1}{n-k} \frac{\hat{\mathcal{E}}'\hat{\mathcal{E}}}{\sigma^2} \right)^{-1}.$$

Note that $C\hat{\beta} - C\beta \mid X \sim N(0, \sigma^2 C(X'X)^{-1}C')$. By the definition of the chi-square distribution, $(C\hat{\beta} - C\beta)'[\sigma^2 C(X'X)^{-1}C']^{-1}(C\hat{\beta} - C\beta) \mid X \sim \chi^2(r)$. By the same argument as before, this is independent of $\hat{\mathcal{E}}'\hat{\mathcal{E}}$ conditional on X. In sum, $F_{0,C} = (\frac{A}{r})(\frac{B}{n-k})^{-1}$ where $A \mid X \sim \chi^2(r)$, $B \mid X \sim \chi^2(n-k)$, and $A \perp B \mid X$. By the definition of the F-distribution, $F_{0,C} \mid X \sim F(r, n-k)$. ■

A joint confidence set for β leads to a uniform confidence band for the population regression line. For each x, it boils down to maximizing or minimizing $x'b$ with respect to b subject to b belonging to the confidence set; for an elliptic confidence set with Theorem 7.8, b belongs to it if and only if $F_C \leq c$ for $\beta = b$ and the critical value c corresponding to the confidence level. Figure 7.3 shows an example of a 99% uniform confidence band for the first regression in Example 7.5 when C is set to be the identity matrix. Note that this is not a 99% prediction band, so the band does not contain 99% of the observations; rather, the probability that this band could have realized at a place that contains the true regression line was 99%.

[5] A possibly more justifiable way to account for the error of asymptotic approximation is to use the Edgeworth expansion or the Berry–Esseen theorem (Proposition 3.6).

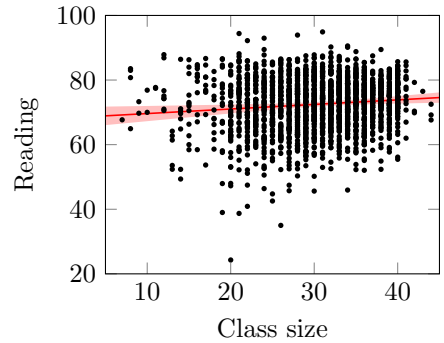


FIGURE 7.3. 99% uniform confidence band around the regression line in Figure 7.9A.

7.2.5. Analysis of Variance (ANOVA). The orthogonality $\mathbb{E}[x\varepsilon] = 0$ implies

$$\mathbb{E}[y^2] = \mathbb{E}[(x'\beta)^2] + \mathbb{E}[\varepsilon^2].$$

If x has an intercept, we may subtract $\mathbb{E}[y]^2$ from both sides and write

$$\mathrm{Var}(y) = \mathrm{Var}(x'\beta) + \mathrm{Var}(\varepsilon).$$

The first term describes the amount of the variance of y that can be attributed to (a linear function of) x, and the second term is the remainder. If $\beta = 0$, then x has no power of explaining y, hence $\mathrm{Var}(y) = \mathrm{Var}(\varepsilon)$; if $x'\beta$ perfectly explains y, then $\mathrm{Var}(y) = \mathrm{Var}(x'\beta)$. In short, the ratio (the "population R^2")

$$\frac{\mathrm{Var}(x'\beta)}{\mathrm{Var}(y)} = 1 - \frac{\mathrm{Var}(\varepsilon)}{\mathrm{Var}(y)}$$

gives the proportion of variation of y explained by $x'\beta$. A plug-in estimator of this,

$$R^2 := 1 - \frac{\frac{1}{n}\sum \hat{\varepsilon}_i^2}{\frac{1}{n}\sum (y_i - \bar{y}_n)^2} = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum (y_i - \bar{y}_n)^2},$$

is called the *coefficient of determination*, or simply *R-squared*, and used as a goodness-of-fit measure. We can also adopt small-sample correction

$$\bar{R}^2 := 1 - \frac{\frac{1}{n-k}\sum \hat{\varepsilon}_i^2}{\frac{1}{n-1}\sum (y_i - \bar{y}_n)^2},$$

which is called the *adjusted coefficient of determination* or *adjusted R-squared*. Note that $\bar{R}^2$ is still a biased estimator of the population R^2; the numerator and the denominator may be unbiased for their own sake, but the reciprocal of an unbiased estimator is not an unbiased estimator of the reciprocal. The R^2 has a nice relation to the customarily reported default F-statistic; $F_{0,C}$ for the hypothesis that all coefficients but for the intercept are zero ($H_0 : \beta_{-0} = 0$) can be written as

$$F_{0,C} = \frac{\frac{1}{k-1}R^2}{\frac{1}{n-k}(1 - R^2)}.$$

When the regressor does not include an intercept, all of the above arguments must be done for the second moments, e.g., $\frac{1}{n}\sum y_i^2$ instead of $\frac{1}{n-1}\sum(y_i-\bar{y}_n)^2$. Then, respective R^2 is called *uncentered*.

Although the above discussion applies to both homo- and heteroskedastic cases, it does not make much sense to decompose the marginal variance of y into the marginal variances of x and ε when heteroskedasticity is concerned. Therefore, R^2 is less used, and the F-statistic—which naturally extends to heteroskedastic errors—is preferred as a primary measure of goodness-of-fit in economics.

7.3. Weighted Least Squares

When sampling is biased or the data is collapsed into a frequency table, a simple average of the dataset fails to estimate the intended population average. In some cases, we can get rid of this bias by weighting.

The *weighted least squares (WLS)* estimator is defined as the solution to

$$\min_{b\in\mathbb{R}^k}\sum_{i=1}^n w_i(y_i - x_i'b)^2$$

for some nonnegative weights w_i. The following are practical situations to use WLS. The first three are to recover β in the population linear projection, while the fourth is to have a smaller variance than OLS under conditional linearity.

(1) *Frequency weighting* ($w_i = n_i$): When observation i represents n_i observations, this weighting duplicates the observation n_i times. This appears, e.g., when categorical observations are converted into frequency tables. This weighting corresponds to `[fweight=`n_i`]` in Stata.
(2) *Analytic weighting* ($w_i = n_i$): When observation i is an average of n_i observations, this weighting inflates the observation by $\sqrt{n_i}$. This comes from the fact that the averaged equation $\bar{y}_i = \bar{x}_i'\beta + \bar{\varepsilon}_i$ shrinks the standard deviation by $1/\sqrt{n_i}$ under independence. This situation occurs, e.g., when observations are anonymized into group-wise averages. This weighting is implemented as `[aweight=`n_i`]` in Stata.
(3) *Inverse probability weighting* ($w_i = p_i^{-1}$): When observation i is sampled with probability p_i, this weighting counterbalances the sampling bias. Dividing by p_i downweights overrepresented observations and upweights underrepresented ones, thereby restoring the sample that properly represents the population. This is used when sampling frequency is uneven, e.g., for stratified sampling. This weighting is specified as `[pweight=`p_i^{-1}`]` in Stata.
(4) *Inverse variance weighting* ($w_i = \mathbb{E}[\varepsilon_i^2 \mid x_i]^{-1}$): When conditional linearity holds in an unsaturated model and conditional heteroskedasticity is known or estimable, this weighting yields a more efficient estimator than the plain vanilla OLS.

The first three are practically important, though not much theory to tell. The fourth is specifically called the *generalized least squares (GLS)* and is related to exploiting conditional linearity to improve efficiency (Remark 7.3).

Proposition 7.10 (Semiparametric efficiency of feasible GLS)**.** *Suppose that* $\mathbb{E}[\varepsilon_i^2 \mid x_i]$ *is bounded away from* 0 *and there is a uniformly consistent estimator* $\widehat{\mathbb{E}[\varepsilon_i^2 \mid x_i]}$ *of* $\mathbb{E}[\varepsilon_i^2 \mid x_i]$ *on the support of* x_i*. Under Assumptions 7.1 and 7.3,*

$$\hat{\beta}_{\text{FGLS}} := \arg\min_{b \in \mathbb{R}^k} \sum_{i=1}^n \widehat{\mathbb{E}[\varepsilon_i^2 \mid x_i]}^{-1} (y_i - x_i' b)^2$$

is semiparametrically efficient and satisfies

$$\sqrt{n}(\hat{\beta}_{\text{FGLS}} - \beta) \rightsquigarrow N\Big(0,\, \mathbb{E}[\mathbb{E}[\varepsilon_i^2 \mid x_i]^{-1} x_i x_i']^{-1}\Big).$$

Proof. Proceed as in Theorem 7.1 and [**vdV98**, Example 25.28]. ■

One drawback of feasible GLS is that if Assumption 7.3 fails, it loses the interpretation as the linear projection of $\mathbb{E}[y \mid x]$ onto x. It is still a linear projection with respect to *some* inner product, but we don't know what it is. A further reading in [**RW17**]. For the finite-sample optimality of GLS, see [**Hay00**, Proposition 1.7].

7.4. Designing the Regression Equation

Econometric theory often assumes the regression equation as given. However, choosing the right equation is a nontrivial and essential step in applied research. In this section, we see some guidance on how to choose an equation.

7.4.1. The intercept. While the theory works with or without the intercept, in most practical situations, it only makes sense to include it. A natural situation in which we should exclude it is when linear regression is used to decompose y into a weighted average of x, such as hedge fund replication [**HL07**] or the cost of capital decomposition [**CP05**]. We may also exclude it when there is a good reason to believe that the linear projection of y on x passes through the origin. For example, the CAPM predicts that

$$r - r_f = \beta(r_m - r_f) + \varepsilon,$$

where r is the return of an asset of interest, r_f the risk-free rate, and r_m the market return. However, even in this case, the intercept is casually included in practice. Indeed, the intercept, if nonzero, has the interpretation as the "alpha" of the asset, so it is still meaningful to examine its nullity rather than assuming it. One possible exception may be that to test the hypothesis that the alpha is zero, we may run the regression without the intercept and choose to do the LM test (Section 6.5), but this is still somewhat contrived. Another example is the panel data model in which the intercept is replaced with fixed effects, but this is rather a generalization of the intercept than a case without it.

The bottom line is that the intercept is included in almost all linear regressions carried out in economics research.

7.4.2. Dummy variables. The regression framework requires that x_i is a numerical variable. Categorical data, therefore, need to be transformed into numerical data. The standard method is to use the *dummy variable* (or *design variable*).

Suppose that $\tilde{x}_i$ is a variable representing the gender, that is, $\tilde{x}_i \in \{\text{male}, \text{female}\}$. To include it in the regression, we create a new variable $x_{i1} = \mathbb{1}\{\tilde{x}_i = \text{male}\}$ and include it in the regression. Suppose we run the regression on $x_i = (1, x_{i1})'$,

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i.$$

By orthogonality, $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[x_{i1}\varepsilon_i] = 0$. This implies $\mathbb{E}[\varepsilon_i \mid x_{i1}] = 0$ and hence

$$\mathbb{E}[y_i \mid x_{i1} = 0] = \beta_0, \qquad \mathbb{E}[y_i \mid x_{i1} = 1] = \beta_0 + \beta_1.$$

Thus, the model satisfies conditional linearity (Assumption 7.3) by construction, and β admits a very clear interpretation. For example, the predictive difference of y_i in gender can be tested by the hypothesis $H_0 : \beta_1 = 0$. If the regression equation automatically satisfies conditional linearity, the model is called *saturated*.

If $\tilde{x}_i$ is a categorical variable that takes k values, then the best practice is to create $k-1$ dummy variables indicating each but one category. For example, let $\tilde{x}_i \in \{0, 1, \ldots, k-1\}$ and define $x_{ij} = \mathbb{1}\{\tilde{x}_i = j\}$. Then the regression equation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{k-1} x_{i,k-1} + \varepsilon_i. \tag{7.1}$$

We can again infer that $\mathbb{E}[\varepsilon_i \mid \tilde{x}_i] = 0$ and hence

$$\mathbb{E}[y_i \mid \tilde{x}_i = 0] = \beta_0, \qquad \mathbb{E}[y_i \mid \tilde{x}_i = j] = \beta_0 + \beta_j$$

for $j = 1, \ldots, k-1$. Again, the model is saturated. The category $\tilde{x}_i = 0$ is called the *base category* and the coefficients on the dummy variables measure the difference of the mean of y_i of their groups relative to the base category.

What happens if we instead run regression on all dummy variables? Write the regression equation as

$$y_i = \alpha + \beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_{k-1} x_{i,k-1} + \varepsilon_i.$$

Since $x_{i0} + \cdots + x_{i,k-1}$ is always 1, we see that the equation

$$y_i = 0 + (\alpha + \beta_0) x_{i0} + (\alpha + \beta_1) x_{i1} + \cdots + (\alpha + \beta_{k-1}) x_{i,k-1} + \varepsilon_i$$

is equally valid. In other words, α and β are not separately identified. This problem is known as *multicollinearity* since the intercept and the sum of x_{ij}s are linearly dependent. Theoretically, it violates invertibility of $\mathbb{E}[x_i x_i']$ and $X'X$.

Alternatively, multicollinearity can be avoided if we exclude the intercept,

$$y_i = \gamma_0 x_{i0} + \gamma_1 x_{i1} + \cdots + \gamma_{k-1} x_{i,k-1} + \varepsilon_i. \tag{7.2}$$

With this formulation, we have $\gamma_j = \mathbb{E}[y_i \mid x_{ij}]$. Therefore, there is a one-to-one relationship with the coefficients in (7.1),

$$\gamma_0 = \beta_0, \qquad \gamma_j = \beta_0 + \beta_j$$

for $j = 1, \ldots, k-1$. This holds in sample as well, so $\hat{\gamma}_0 = \hat{\beta}_0$ and $\hat{\gamma}_j = \hat{\beta}_0 + \hat{\beta}_j$ numerically. Therefore, there is no material difference between (7.1) and (7.2). However, (7.1) is practically preferred since it is easier to handle when there are

other variables to include. If there is more than one categorical variable, having an intercept and excluding one base category from each categorical variable lifts our burden of having to keep track of which variables to include.[6]

A bad practice is linear-in-category modeling. Sometimes a categorical variable is given in the form of numerical values; for example, the Standard Industry Classification (SIC) classifies industries with a 4-digit number. For a numerical categorical variable $\tilde{x}_i$, it might be tempting to run

$$y_i = \beta_0 + \beta_1 \tilde{x}_i + \varepsilon. \tag{7.3}$$

This assumes that the increment of the predicted y_i from $\tilde{x}_i = 0$ to $\tilde{x}_i = 1$ is the same as that from $\tilde{x}_i = 1$ to $\tilde{x}_i = 2$, which does not make sense if $\tilde{x}_i$ is nominal. When $\tilde{x}_i$ is ordinal, there might be cases where this is reasonable, but even so, the dummy variable approach is usually preferred. Note that the dummy variables model subsumes the linear-in-category model, that is, if $\beta_j - \beta_{j-1}$ is constant, (7.1) reduces to (7.3). Therefore, the dummy variables model allows us to estimate the model without imposing (but not excluding) the constant increment case. Also, economists have a general preference toward saturated models. Saturated models grant the interpretation as the conditional expectation without imposing any assumption (other than the existence of moments). In this sense, they are examples of a nonparametric model that is free of our speculation over specification.

7.4.3. Interaction terms. Suppose we want to investigate gender discrimination in salary. We have three variables: salary y_i, gender $\tilde{x}_{i1} \in \{\text{male}, \text{female}\}$, and experience $\tilde{x}_{i2} \in \{\text{experienced}, \text{inexperienced}\}$. Since we have two categorical variables, we may model it as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

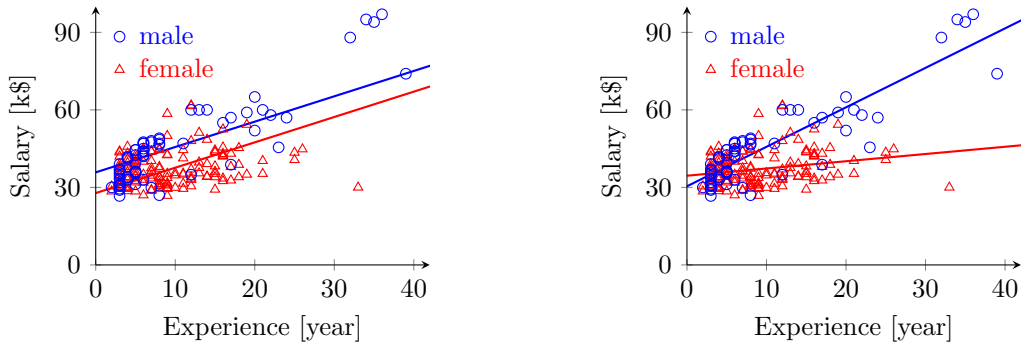
where $x_{i1} = \mathbb{1}\{\tilde{x}_{i1} = \text{male}\}$ and $x_{i2} = \mathbb{1}\{\tilde{x}_{i2} = \text{experienced}\}$. By orthogonality, $\mathbb{E}[\varepsilon_i] = \mathbb{E}[x_{i1}\varepsilon_i] = \mathbb{E}[x_{i2}\varepsilon_i] = 0$. This implies $\mathbb{E}[\varepsilon_i \mid x_{i1}] = \mathbb{E}[\varepsilon_i \mid x_{i2}] = 0$ but *not* $\mathbb{E}[\varepsilon_i \mid x_{i1}, x_{i2}] = 0$. Therefore, the model is not saturated and conditional linearity may not hold. Specifically, this model assumes that the gender difference in salary for experienced employees is the same as the gender difference in salary for inexperienced employees. If male and female are rewarded differently on their experiences, this model cannot capture that aspect. However, it is sensible to expect this as the starting salary is difficult to discriminate.

To capture the gender difference in the reward on experience, consider

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i. \tag{7.4}$$

Then the reward on experience for female is captured by β_2 while that for male by $\beta_2 + \beta_3$. Thus, β_3 captures the gender difference in how experience is rewarded. The newly introduced term $x_{i1}x_{i2}$ is called the *interaction term* between x_{i1} and x_{i2}. Note

[6] An exception is the fixed effect model, in which the intercept is replaced with a set of dummy variables for various groups.



(A) $Salary_i = \beta_0 + \beta_1 Male_i + \beta_2 Exp_i + \varepsilon_i$. This specification sets different intercepts but a common slope across genders.

(B) $Salary_i = \beta_0 + \beta_1 Male_i + \beta_2 Exp_i + \beta_3 Male_i \times Exp_i + \varepsilon_i$. This specification sets different intercepts and different slopes across genders.

FIGURE 7.4. The effect of an interaction term. The model (A) has no interaction term and fits two parallel lines for salary vs. experience that have different intercepts. The model (B) has an interaction term and fits two lines that have different slopes as well as different intercepts.

that orthogonality also implies $\mathbb{E}[x_{i1}x_{i2}\varepsilon_i] = 0$, so we get $\mathbb{E}[\varepsilon_i \mid x_{i1}, x_{i2}] = 0$. In other words, the model is saturated such that

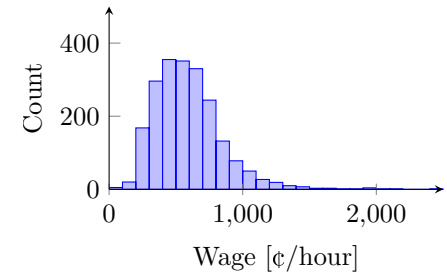
$$\mathbb{E}[y_i \mid x_{i1} = x_{i2} = 0] = \beta_0, \qquad \mathbb{E}[y_i \mid x_{i1} = x_{i2} = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3,$$
$$\mathbb{E}[y_i \mid x_{i1} = 1, x_{i2} = 0] = \beta_0 + \beta_1, \qquad \mathbb{E}[y_i \mid x_{i1} = 0, x_{i2} = 1] = \beta_0 + \beta_2.$$

We can also create interaction terms between a categorical variable and a numerical variable, or between two numerical variables. Suppose that the experience was measured by how many years an employee has worked. Then, β_3 in (7.4) measures the change in the slope of the reward on experience by gender (Figure 7.4). In that case, however, the model is not saturated.

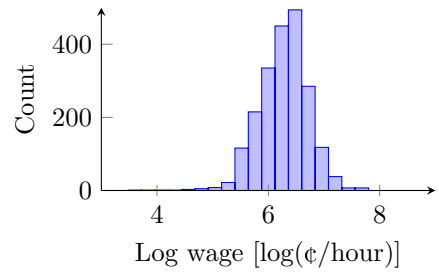
EXERCISE 7.6. Design a regression equation that fits two distinct quadratic curves to salary against experience across genders in Figure 7.4.

7.4.4. Higher-order terms and nonlinear transformations. If x_i is a numerical variable, there is usually little reason to expect that the conditional expectation of y_i given x_i is linear in x_i. However, linear regression is not constrained to models that are linear in variables. Adding polynomials such as x_i^2 and x_i^3 is a powerful way to make linear regression flexible. These are called *higher-order terms*. If there is more than one variable, then we may consider adding interactions of higher-order terms as well. In a way, x_i^2 is an interaction term of x_i with itself.

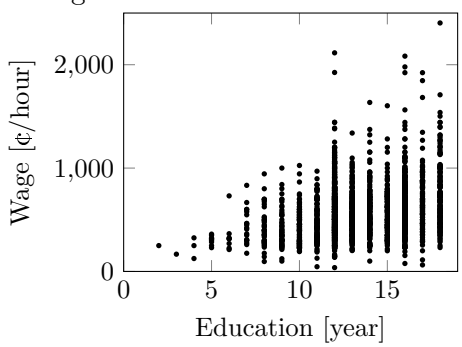
We may also use other transformations. In the context of time series, trigonometric functions are used to account for seasonality, known as *harmonic regression*. When we suspect that the effect of x_i on y_i is multiplicative, the logarithm function is used. For example, the effect of schooling on wages is considered multiplicative in a sense that receiving one more year of education would multiply—rather than add—to



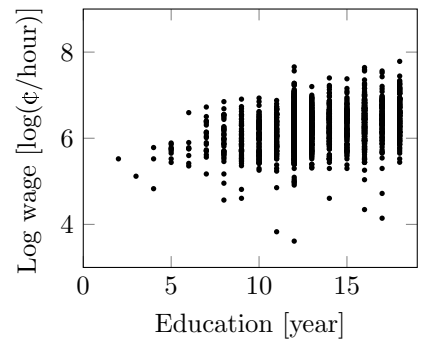
(A) Wage distribution is asymmetric and looks lognormal.



(B) Log wage distribution looks more symmetric and normal.



(C) Wage against schooling shows nonlinearity and heteroskedasticity.



(D) Log wage against schooling shows less nonlinearity and heteroskedasticity.

FIGURE 7.5. Logarithm transformation of wages. Data from [**Car95**].

one's current wage (Figure 7.5). Therefore, the *Mincer earnings regression* considers

$$\log wage_i = \beta_0 + \beta_1 schooling_i + \varepsilon_i.$$

Since wages are nonnegative and the income distribution resembles lognormal, this specification makes sense.[7] The log transformation is also used for time series data to deal with the rate of change of a variable. For example, when we consider the process of stock returns, we may regress the log stock price on its past values as

$$\log p_t = \beta_0 + \beta_1 \log p_{t-1} + \varepsilon_t.$$

While we could have explicitly computed the stock return by $\frac{p_t - p_{t-1}}{p_{t-1}}$, the logarithm emulates the similar effect since

$$\frac{p_t - p_{t-1}}{p_{t-1}} \approx \log\left(1 + \frac{p_t - p_{t-1}}{p_{t-1}}\right) = \log p_t - \log p_{t-1},$$

and it does not shave off the sample size by one. If the price is considered in continuous time, the log price makes more sense in that it represents an infinitesimal return.

There are also situations in which economic theory suggests a specific regression equation. Suppose we are interested in how the demand Q changes in response to the

[7] If income is determined by the product of independent factors, the CLT implies that it distributes according to a lognormal [**HS15**].

change in price P. If we model it as

$$Q_i = \beta_0 + \beta_1 P_i + \varepsilon_i,$$

we are assuming that the effect of the price increase from \$100 to \$110 is the same as that from \$1,000 to \$1,010, and that the effect on the goods . However, it is rather sensible to assume that the change from \$100 to \$110 is more comparable with the change from \$1,000 to \$1,100. Similarly on the quantity side, it makes more sense to think in terms of the relative change rather than the absolute change so that the effect is measured in units independent of the market size. In this case, the effect of price on quantity demanded we care about is not $\frac{\Delta Q}{\Delta P}$ but rather $\frac{\Delta Q/Q}{\Delta P_i/P}$. For infinitesimal changes, this is equivalent to $\frac{\Delta \log Q}{\Delta \log P}$, which is called the *price elasticity of demand*. This motivates the regression equation

$$\log Q_i = \beta_0 + \beta_1 \log P_i + \varepsilon_i.$$

On the supply side, the functional form of a production function that is analytically the easiest to handle is known as the *Cobb–Douglas production function* and is given by $Y = AK^\alpha L^\beta$ where Y is the production, K the capital, L the labor, and A the total factor productivity. This motivates the regression

$$\log Y_i = \gamma + \alpha \log K_i + \beta \log L_i + \varepsilon_i,$$

where α and β have the interpretation as the output elasticities of capital and labor, and γ the logarithm of the total factor productivity. In macroeconomics, the Solow growth model motivates a specific type of regression equations for investigating economic growth [**Ace09**, Section 3.2].

Example 7.3 (Accounting conservatism). Accounting practice has long employed the principle of conservatism that "anticipates no profits but anticipates all losses." At first sight, one might suspect that this is a company's effort to avoid corporate taxes, but such conservatism predates corporate taxes, shareholder litigation, and accounting regulation, so the origin is actually not clear. In the literature, there is an explanation of conservatism as a commitment device to cope with information asymmetry between managers and claimholders. [**Bas97**] considers an asymmetric regression to estimate the degree of accounting conservatism. If the market is efficient, the stock return reflects all public information and hence is a good proxy for fair economic value. This motivates the following equation

$$I_{it} = \alpha_0 + \alpha_1 \mathbb{1}\{R_{it} < 0\} + \beta_0 R_{it} + \beta_1 R_{it} \mathbb{1}\{R_{it} < 0\} + \varepsilon_{it},$$

where I_{it} and R_{it} are the earnings and stock return for firm i in year t. Here, α_1 and β_1 measure the changes of the intercept and slope in the correspondence between the accounting earnings and stock returns when the market returns are negative.

7.4.5. Residual diagnostics. So far, we had some theoretical guidance on the design of the regression equation. In many cases, however, there is only so much theory can tell. In general, how can we "detect" wrong specification when we want to recover the conditional expectation?

If there is only one regressor, scatterplotting y_i against x_i reveals a rough shape of $\mathbb{E}[y_i \mid x_i]$ (as in Anscombe's quartet). When there is more than one regressor, plotting a multidimensional relationship is usually hard, so we need to rely on some dimension reduction before visualizing the data. However, scatterplotting y_i against each regressor does not provide much information since their relationship is heavily affected by the variation of other regressors. To circumvent this problem, we often plot the residual $\hat{\varepsilon}_i$ against each regressor. Such heuristic visual assessment is called *residual diagnostics*.

If the conditional linearity (Assumption 7.3) holds, we have $\mathbb{E}[\varepsilon_i \mid x_i] = 0$, which further implies $\mathbb{E}[\varepsilon_i \mid x_{ij}] = 0$ for every j by the law of iterated expectations (Theorem 2.9). Therefore, if the scatter plot of $\hat{\varepsilon}_i$ against some regressor exhibits nonlinearity, we suspect that the conditional expectation function is different from the specification of the linear regression. Meanwhile, having $\mathbb{E}[\varepsilon_i \mid x_{ij}] = 0$ for every j does not imply $\mathbb{E}[\varepsilon_i \mid x_i] = 0$, so there are specification errors that stay under the radar of these plots. One possible remedy for this is to use the fact that $\mathbb{E}[\varepsilon_i \mid x_i] = 0$ if and only if $\mathbb{E}[\varepsilon_i \mid f(x_i)] = 0$ for every real-valued measurable function $f : \mathbb{R}^k \to \mathbb{R}$. Thence, we can plot the residual against various one-dimensional transformations of the regressors (e.g., the fitted value) to detect misspecification in finer detail.

These plots are also useful in detecting a possible violation of conditional homoskedasticity (Assumption 7.4). Also, when we suspect the data might contain outliers, these plots help discern outliers in ε_i from simply large observations of ε_i due to conditional heteroskedasticity.

Figure 7.6 shows an example of the residual diagnostics for the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i.$$

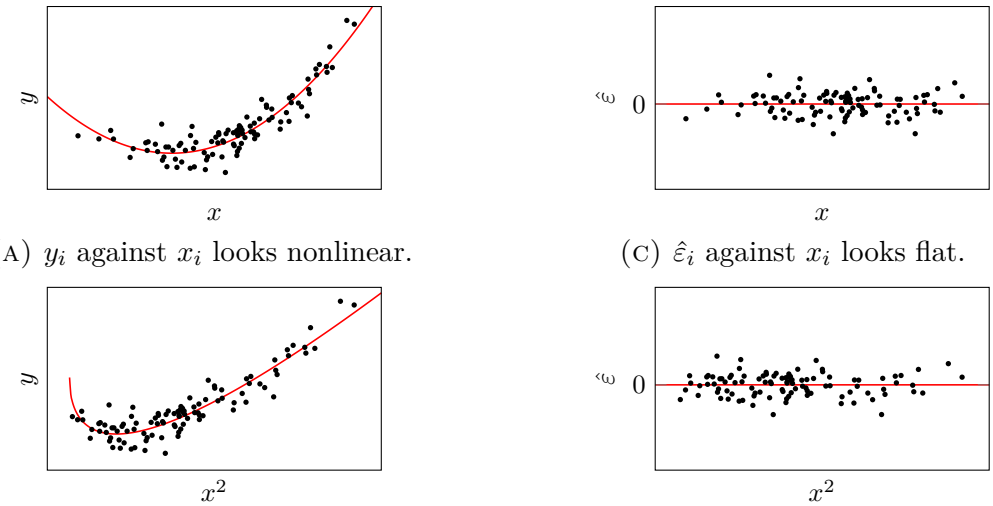
The scatterplots of y_i against x_i and x_i^2 exhibit inherent nonlinearity and are not helpful in judging the specification accuracy (Figures 7.6A and 7.6B). Meanwhile, the plots of $\hat{\varepsilon}_i$ against x_i and x_i^2 show flat overall locations as well as constant variations (Figures 7.6C and 7.6D), so we can diagnose that the conditional linearity and conditional homoskedasticity hold under this specification and that there are no obvious outliers.

7.5. Specification Search

Given the flexibility of linear regression, it is tempting to engage in specification searching. That is, for a given dataset, we try various regressions to hunt for the "best" one, be it the highest R^2 or the most significant $\hat{\beta}$. There are two problems associated with this practice: (1) *overfitting* and (2) *data snooping*, or more cynically, *p-hacking*.[8]

The problem of overfitting occurs when we include too many regressors. Suppose that y is independent of all regressors. Then the population equation $y = x'\beta + \varepsilon$ holds with $\beta = 0$, meaning $\mathrm{Var}(x'\beta) = 0$ and $\mathrm{Var}(\varepsilon) = \mathrm{Var}(y)$. However, if we include as many regressors as observations, the residual can be made identically zero and $x'\hat{\beta}$

[8]Not to be confused with *data mining*, which is a fine branch of statistics and computer science.



(A) y_i against x_i looks nonlinear.

(C) $\hat{\varepsilon}_i$ against x_i looks flat.

(B) y_i against x_i^2 looks nonlinear.

(D) $\hat{\varepsilon}_i$ against x_i^2 looks flat.

FIGURE 7.6. Residual diagnostics for $\mathbb{E}[y_i \mid x_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$. The plots (C) and (D) indicate that the conditional expectation seems correctly specified. Meanwhile, the plots (A) and (B) are not very helpful in diagnosing misspecification.

seems to explain the entirety of y. If we use small-sample correction (Proposition 7.7), we get a warning that the standard error is undefined rather than zero, but this is not a fundamental solution to the problem of overfitting.[9]

A practical solution to overfitting is to use methods that can account for it. For example, the *least absolute shrinkage and selection operator (LASSO)* method selects regressors from a large pool when only a handful of them are strong predictors [**BC11**]. The *ridge regression* also makes regression possible when there are more regressors than the sample size.

A similar but distinct situation is when we have many statistical models to try, instead of many regressors to try in one regression model. A general solution to compare different models is to appeal to sample splitting techniques such as the *holdout method*, *cross-validation*, and *cross-fitting*. If the models admit a common structure, there may be a metric by which to compare them, such as the *Akaike information criterion (AIC)* and the *Bayesian information criterion (BIC)* [**Gre18**, Section 5.8].

Meanwhile, the problem of p-hacking is much nastier. This occurs when we run several regressions to find a significant coefficient and hide insignificant results. As stated in Section 5.6, this practice is problematic since the statistical discovery is upheld by the rarity of a rejection. If we are allowed to observe many estimates, we would casually encounter a "rare" one by chance; then, cherry-picking the estimates for reporting makes rarity common.

[9]Also, Bessel-type correction is often not available in nonlinear models.

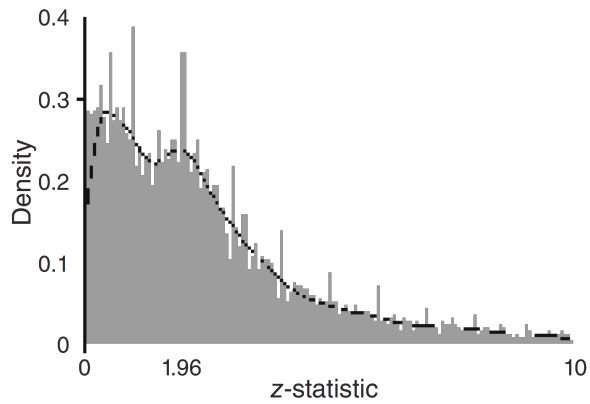


FIGURE 7.7. Distribution of t-statistics in published articles has a hump around 1.96 [**BLSZ16**, Figure 1].

In practice, there are some situations in which a researcher might be motivated to do so. For example, a dataset might be obtained with huge costs, and the researcher might be inclined to write a few papers off of it even after the researcher's initially intended results could not have been obtained; a journal may have a tendency to publish significant results more often than insignificant ones (*publication bias*, Figure 7.7), and the researcher may not care to report a vast number of insignificant results generated in the course of research. However, reporting a selected few as if they are a legitimate discovery (i.e., as if they are the only ones tested) is a form of fabrication of research. In many cases, such mistakes are made without bad intentions but with the lack of care or unfamiliarity with statistics.

This vulnerability to selective reporting is inherent in any scientific discipline whose primary mode of reasoning is inductive and exact reproduction of the data is costly, thus notably in social science and medicine. In these fields, it is sometimes reported that a number of previous findings cannot be replicated in the followup research. This is often provocatively called the *replication crisis*.

To cope with this problem, these disciplines have adopted a few changes over the past decades. The first is *pre-analysis plan*. There are repositories that store plans of research, including what data to collect, which regression equations to estimate, and what tests to examine. The analysis plans should be submitted *before* the data is collected. With this, researchers have a way to prove that what they report in the paper is not a result of p-hacking but a legitimate discovery. If they want to try more specifications, they can do so with a specific note clarifying that the additional results are obtained after they observed the data, leaving the readers the discretion to take them with a grain of salt.

The second change is that more and more journals are accepting results that are not statistically significant. If the research is well-motivated and the method is sound, the fact that few significant results are found is not necessarily seen as a downside or a lack of contribution. Along with this, some journals now prohibit authors to put a star "$*$" next to the estimates in the regression table that indicates significance, a

practice that once was so popular that it endorsed screening papers based thereon without much attentive reading.

Statistics also provides some methods to guard against p-hacking, albeit the use of them still requires discipline and honesty of the users. For example, multiple hypothesis testing provides a way to take into account the dilution of rarity when many hypotheses are tested, in which the proportion of rejections instead of the probability of each rejection is controlled. See Section 5.6 and references therein. Recently, a statistical method to test the existence of p-hacking was also proposed, and gave a mixed conclusion when applied to economics papers [**EKW22**].

7.6. Simpson's Paradox and the Frisch–Waugh Theorem

Does adding a new regressor change the coefficient of an existing regressor? The answer is not only affirmative, but the change can sometimes be drastic and counter-intuitive. The phenomenon that a coefficient of a regressor flips its sign after adding a new variable is known as *Simpson's paradox*.

Example 7.4 (Graduate admission at Berkeley). In 1973, the associate dean of the graduate school at UC Berkeley found that only 35% of female applicants were admitted compared to 44% of male applicants, and asked Peter Bickel, a renowned professor of statistics, to investigate the issue [**BHO75**]. The sample consists of applications to the graduate programs in the six largest majors [**FPP07**, Section 2.4]. The regression of admission on gender gives

$$\widehat{Admitted}_i = \underset{(0.011)}{0.30} + \underset{(0.014)}{0.14}\, Male_i,$$

where the numbers in parentheses are the robust standard errors. This roughly means that male applicants are admitted 14% more than female applicants, and this coefficient is statistically significant. When we include the dummies for majors, we get

$$\widehat{Admitted}_i = \underset{(0.020)}{0.66} - \underset{(0.015)}{0.02}\, Male_i - \underset{(0.025)}{0.01}\, B_i - \underset{(0.023)}{0.30}\, C_i - \underset{(0.023)}{0.31}\, D_i - \underset{(0.025)}{0.40}\, E_i - \underset{(0.019)}{0.59}\, F_i.$$

In this regression, the coefficient on gender flips its sign and is no longer significant. One explanation of this is the difference in the numbers of applicants across majors. Figure 7.8 shows that many male applicants applied to less selective majors, like mechanical engineering, and many female applicants to more selective majors, like English.

Example 7.5 (Class size and test scores). The relationship between class size and education quality is not trivial. One may speculate that the smaller the class size, the more time and attention a teacher can spend for each student, hence the higher educational effect; however, the observed relationship is not always so straightforward. Using the data from Israeli public schools, [**AL99**, Table II] report that regressing 4th graders' test scores for reading comprehension on class size gives

$$\widehat{Reading}_i = \underset{(1.12)}{68.2} + \underset{(0.033)}{0.141}\, ClassSize_i.$$

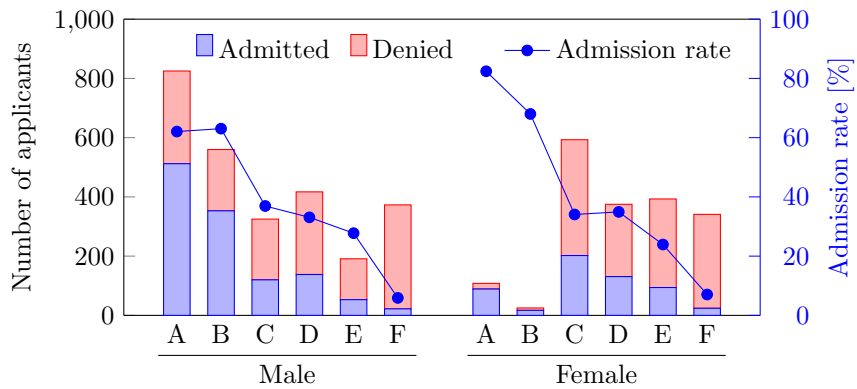


FIGURE 7.8. Berkeley's admission data illustrate Simpson's paradox [**BHO75**]. Six majors are labeled through A to F. The overall admission rate in the six majors for males is 44.5% while for females is 30.4%, exhibiting a huge difference. However, the admission rate conditional on each major does not seem to differ much by gender. A large factor contributing to the lopsided total admission rates is the fact that many male applicants have applied to less selective majors (A and B) while many female applicants to more selective ones (C to F).

Therefore, the predicted test score *decreases* with a reduced class size (Figure 7.9A). When we include the percentage of students with disadvantaged backgrounds, we get

$$\widehat{Reading}_i = \underset{(1.03)}{78.8} - \underset{(0.028)}{0.053}\, ClassSize_i - \underset{(0.013)}{0.339}\, Disadvantaged_i.$$

This means that the predicted test score *increases* as the class size is reduced, once we control for the percentage of disadvantaged students (Figure 7.9B). Possible reasons may be that larger schools are located in big cities while smaller schools in poor towns where the overall environment is disadvantaged, or that school principals may group lagging students into smaller classes. The authors then estimate the causal effect of a reduced class size using an instrumental variable method (Example 9.15).

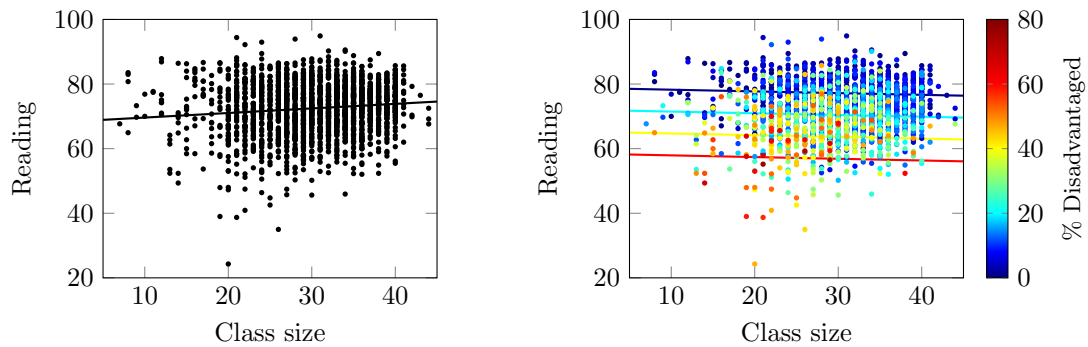
The simplest way to illustrate this phenomenon is to think about adding an intercept. Consider two regression equations

$$y_i = \gamma_1 x_{i1} + \varepsilon_i \qquad \text{versus} \qquad y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i.$$

Demeaning both sides of the second equation gives

$$y_i - \mathbb{E}[y_i] = \beta_1(x_{i1} - \mathbb{E}[x_{i1}]) + \varepsilon_i - \mathbb{E}[\varepsilon_i] = \beta_1(x_{i1} - \mathbb{E}[x_{i1}]) + \varepsilon_i,$$

where $\mathbb{E}[\varepsilon_i] = 0$ by orthogonality. Thus, the regression line for the second equation goes through $(\mathbb{E}[x_{i1}], \mathbb{E}[y_i])$ while that for the first through $(0, 0)$. In other words, the first regression imposes the restriction that the linear line passes through the origin, while the second does not and *finds out* that the best adjustment of the level is such that the line passes through $(\mathbb{E}[x_{i1}], \mathbb{E}[y_i])$.



(A) Regressing the reading score on class size finds a significantly positive slope.

(B) Controlling for the percentage of disadvantaged students makes the coefficient on class size negative.

FIGURE 7.9. Simpson's paradox in the relationship of the average reading score and the class size for 4th graders [**AL99**, Table II (7)–(8)]. When conditioned on the percentage of students with disadvantaged backgrounds, the slope of class size on reading flips its sign.

This exercise also reveals that β_1 is equivalent to γ_1 where both y_i and x_{i1} are replaced by their demeaned versions. Mathematically,

$$\gamma_1 = \frac{\mathbb{E}[x_{i1}y_i]}{\mathbb{E}[x_{i1}^2]}, \qquad \beta_1 = \frac{\mathrm{Cov}(x_{i1}, y_i)}{\mathrm{Var}(x_{i1})}.$$

EXERCISE 7.7. Show that β_1 is equivalent to γ_1 where x_{i1} is demeaned but not y_i.

If adding an intercept creates an effect of demeaning variables, what effect does adding a random variable create? In linear projection terms, demeaning can be considered the projection onto the orthocomplement of the space spanned by the intercept—the demeaned variable and the intercept have a zero cross moment. This observation generalizes to other random variables. Consider

$$y_i = x_{i1}'\beta_1 + x_{i2}'\beta_2 + \varepsilon_i.$$

(The intercept may or may not be in either vector x_{i1} or x_{i2}.) Let M_2 be the projection operator onto the orthocomplement of x_{i2}, that is,

$$M_2(z) \coloneqq z - x_{i2}'\mathbb{E}[x_{i2}x_{i2}']^{-1}\mathbb{E}[x_{i2}z].$$

Then, β_1 can be found in the regression

$$M_2(y_i) = M_2(x_{i1}')\beta_1 + \varepsilon_i.$$

This argument goes through just as well in the sample (by replacing expectations with sample averages and parameters with estimates), which is known as the *Frisch–Waugh theorem*.

Theorem 7.11 (Frisch–Waugh)**.** *Let $Y = X\hat{\beta} + \hat{\mathcal{E}} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\mathcal{E}}$ be an estimated equation. Suppose that $X'X$ is invertible. Then, $\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y$ where $M_2 \coloneqq I - X_2(X_2'X_2)^{-1}X_2'$.*

PROOF. Since $M_2X_2 = 0$ and $X_2'\hat{\mathcal{E}} = 0$, multiplying $X_1'M_2$ to $Y = X\hat{\beta} + \hat{\mathcal{E}}$ yields $X_1'M_2Y = X_1'M_2X_1\hat{\beta}_1 + X_1'\hat{\mathcal{E}} = X_1'M_2X_1\hat{\beta}_1$. Then $X_1'M_2X_1$ is invertible if $X'X$ is. ■

This theorem tells us that including X_2 does not change $\hat{\beta}_1$ if X_2 is orthogonal to X_1, that is, if the sample correlation of X_1 and X_2 is zero. In population terms, if the correlation of x_{i1} and x_{i2} is zero, the target coefficient on x_{i1} remains unchanged.

Note that the projection onto the orthocomplement of X_2 is nothing but the residual of the regression on X_2. Therefore, we can estimate β_1 by running two smaller regressions. Let the regression of X_1 on X_2 be

$$X_1 = X_2\hat{\Pi} + \hat{V}.$$

Since $\hat{\Pi} = (X_2'X_2)^{-1}X_2'X_1$, we get $\hat{V} = X_1 - X_2(X_2'X_2)^{-1}X_2'X_1 = M_2X_1$. This means that in order to obtain the coefficient of X_1 in the regression of Y on (X_1, X_2), we can first regress X_1 on X_2 and then regress Y on the residuals of the first regression.[10] This elimination of X_2 from the original regression is called *partialing out*, and historically was used to reduce inversion of a huge matrix into inversion of smaller matrices in the old days when computers were not as powerful. Today, partialing out has no practical value, but it still retains theoretical value in simplifying the analysis of regression and in understanding the problem of endogeneity for causal inference.

7.7. Nonparametric Regression

The model cannot be saturated if a regressor is continuous. However, we can saturate the model asymptotically if $\mathbb{E}[y \mid x]$ is smooth. For example, if $\mathbb{E}[y \mid x]$ is analytic, then it has a representation by a Taylor series. Then, if we gradually add higher-order polynomials as n gets large, we might be able to recover $\mathbb{E}[y \mid x]$ in the limit. The *polynomial regression* considers

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i,$$

where k is set to diverge at an appropriate rate as $n \to \infty$. If $\mathbb{E}[y \mid x]$ is a square-integrable function on a bounded domain, it has a representation by a Fourier series. Then, by gradually adding higher-frequency trigonometric functions, we might recover $\mathbb{E}[y \mid x]$ in the limit. These are examples of an adaptive estimation method called the *sieve estimation*, which involves optimization over a gradually enriched sequence of models.[11]

Generally, there are many different methods to recover the conditional expectation nonparametrically. They can be largely classified into two categories: the global approach that aims to approximate $\mathbb{E}[y \mid x]$ as a whole function and the local approach that aims to approximate $\mathbb{E}[y \mid x]$ at each point or on a small region.

(1) *Global linear approach.* The method that uses a sequence of linear regression models that gradually adds more and more regressors. It is called *series estimation* and includes *polynomial regression* and *harmonic regression.*

[10] There is no need to regress Y on X_2 since M_2 is idempotent.

[11] "Sieve" is pronounced /sív/, not /sí:v/.

(2) *Global nonlinear approach.* The method that uses an expanding (usually nested) sequence of nonlinear regression models that become more and more flexible. For example, *neural networks* and *random forests* fall into this category.
(3) *Local approach.* The method that uses only some neighboring observations to approximate $\mathbb{E}[y \mid x]$ locally. Examples include *local linear regression*, *spline regression*, *kernel regression*, and *k-nearest neighbor regression*.

In all of these methods, it is crucial to avoid overfitting in finite samples while ensuring that the model saturates asymptotically. The parameters that control the flexibility of the model are called the *tuning parameters* and should be carefully adjusted as the sample size increases. Some methods have a data-driven way of picking the tuning parameters, but there is always some degree of arbitrariness in their choices. The problem of fine-tuning is especially pronounced in advanced machine learning methods.

In causal inference, nonparametric regression is not used as often as it can be. Part of the reason may be a communication issue. If we have many conditioning variables, it is not obvious how to best present a nonlinear function of them and what to conclude with it. In contrast, parsimonious linear regression can be reported as a familiar table and usually gives one number $\hat{\beta}_1$ that is the most relevant in drawing conclusions from the analysis. On the other hand, if prediction is concerned, there is no issue related to interpretability, so nonparametric regression is a popular and powerful choice.

7.A. Navigating Through the Regression Tables

To illustrate the practical flow of linear regression, let us take Example 7.4. The data is given as a frequency table in Table 7.1. To run regression, we first transform this into 24 observations with variables $Admitted_i$, $Male_i$, $Major_i$, and n_i, where n_i represents the number of applicants for each major and gender. This is still a collapsed dataset, so we use the WLS with frequency weighting (Section 7.3). In Stata, the regression (second specification) is done with the command

```
reg admit male i.maj [fweight=n], vce(r)
```

where the option `vce(r)` specifies the heteroskedasticity-robust standard error with small-sample correction ($\hat{V}$ in Section 7.2.4).

Table 7.2 is the regression output. The upper right corner has five numbers. `Number of obs` is the sample size n. `F(6, 4519)` is the F-statistic for the joint hypothesis that all but the intercept coefficients are zero; the numbers 6 and 4519 are the degrees of freedom, r and $n-k$. `Prob > F` is the p-value for the same hypothesis, using the F distribution; the fact that this is small indicates that the regressors have strong predictive abilities for y. `R-squared` is the (unadjusted) R^2; this is small, so there is still much variation of y that is not explained by the regressors.[12] `Root MSE`

[12] A terrible goodness-of-fit measure is a common observation for binary dependent variables.

TABLE 7.1. Admission data for Example 7.4 [**FPP07**, p. 18].

	Male		Female	
Major	Number of applicants	Percent admitted	Number of applicants	Percent admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

TABLE 7.2. Stata's `reg` output for the second regression in Example 7.4.

```
Linear regression                               Number of obs     =      4,526
                                                F(6, 4519)        =     240.47
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1724
                                                Root MSE          =     .44361

------------------------------------------------------------------------------
             |               Robust
       admit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -.0184252   .0147571    -1.25   0.212    -.0473563    .0105059
             |
         maj |
           B |  -.0103346   .0254036    -0.41   0.684    -.0601381    .0394689
           C |  -.3031654   .0234421   -12.93   0.000    -.3491234   -.2572074
           D |  -.3111034   .0234469   -13.27   0.000    -.3570708    -.265136
           E |  -.4027126   .0251705   -16.00   0.000    -.4520591   -.3533661
           F |  -.5863997   .0187201   -31.32   0.000    -.6231002   -.5496993
             |
       _cons |    .660451   .0200397    32.96   0.000     .6211634    .6997386
------------------------------------------------------------------------------
```

is the square root of $\widehat{\mathbb{E}[\varepsilon_i^2]} = \frac{1}{n-k}\sum \hat{\varepsilon}_i^2$; this shows the unexplained variation in the units of y while R^2 is a unitless measure.

In the table below, the first column lists the dependent variable and the regressors; `_cons` is the intercept term. The second column shows each estimate of β_j. The third column shows the heteroskedasticity-robust marginal standard error of each $\hat{\beta}_j$. The fourth column gives the t-statistic for the hypothesis that each β_j is zero; this is simply the estimate divided by the standard error. The fifth column gives the p-value for this hypothesis, using the t distribution. The sixth and seventh columns provide the

TABLE 7.3. Regression estimates of two specifications in Example 7.4.

	(1) Admitted	(2) Admitted
Male	0.142***	−0.0184
	(0.0144)	(0.0148)
Major B		−0.0103
		(0.0254)
Major C		−0.303***
		(0.0234)
Major D		−0.311***
		(0.0234)
Major E		−0.403***
		(0.0252)
Major F		−0.586***
		(0.0187)
Constant	0.304***	0.660***
	(0.0107)	(0.0200)
N	4,526	4,526
$\bar{R}^2$	0.020	0.171
F	96.89	240.5

Heteroskedasticity-robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. There are six majors labeled through A to F, and the base category is chosen to be A.

95% confidence interval for each coefficient. Note that the fourth to seventh columns can be recovered from the knowledge of the second and third columns.

Table 7.3 is an example of the regression table often shown in a paper. The first column usually lists the regressors; at the bottom, various other information such as the sample size (typically denoted by N), the adjusted R^2, or the F statistic is provided. The second and third columns respectively constitute separate regression specifications. The second column, labeled as (1), uses only the intercept and gender as the regressors, while the third column, (2), includes the major dummies. The stars next to the coefficients are explained at the bottom of the table; for this example, * indicates significance at 5%, ** at 1%, and *** at 0.1%. Be careful that more and more journals are starting to employ the policy that the stars should *not* be placed in the table. The numbers in the parentheses below the estimates are the robust standard errors; in some cases, the t-statistics or the p-values are shown in the parentheses. Some authors do not list all the regressors in the table, especially when the regression equation contains hundreds of regressors.

CHAPTER 8

Logistic Regression

Choose well. Your choice is brief, and yet endless.

THE MASON LODGE, JOHANN VON GOETHE, TRANSLATED BY THOMAS CARLYLE, 1827

When the dependent variable is categorical, a direct application of some regression methods may not make much sense. The methods that are specifically designed to handle this case are called *classification*. In economics, this situation arises frequently when we want to explain how humans make decisions. The decision of "whether to buy something" or "which school to attend" is of categorical nature, and we aim to explain it with a variety of economic observations such as gender, age, and income. The "whether" problems are called the *binary choice models* (or the *binary regression*) and "which" problems are called the *multiple choice models* (or the *multinomial regression*). They are collectively called the *discrete choice models*. The word "choice" may be replaced with "response."

Note that the classification problem is not any different from the regression problem when the set of regressors is also discrete. For example, when we want to explain a decision of purchasing seasonings y_i by gender x_i, we can saturate the model with

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

without any loss of generality. Here, $\beta_0 = P(y_i = 1 \mid x_i = 0)$ and $\beta_1 - \beta_0 = P(y_i = 1 \mid x_i = 1)$, representing a complete description of the conditional distribution of y_i given x_i.

The linear form becomes problematic when the model is not saturated. For example, if we want to explain the purchase y_i by income z_i, the regression

$$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$$

is not without consequences. Is it reasonable to expect a linear relation between a bounded variable y_i and an unbounded variable z_i? Classification methods deal with this inherent nonlinearity.

8.1. Logistics of the Logistic Regression

The *logistic regression* is one of the most basic classification methods that has nice analogies with linear regression. When the dependent variable is binary, it is called the *binomial logistic regression*; when the dependent variable takes more than two

(unordered) categories, the *multinomial logistic regression*. Let us first consider the binary case. The logistic regression specifies the relation between y and x as

$$\log \frac{P(y=1 \mid x)}{P(y=0 \mid x)} = \log \frac{P(y=1 \mid x)}{1 - P(y=1 \mid x)} = x'\beta.$$

The left-hand side is called the *log-odds ratio*, and the function $p \mapsto \log \frac{p}{1-p}$ is called the *logit function*. The log-odds ratio is a way to map the probability in the scale of $(0,1)$ to a real number in the scale of $(-\infty, \infty)$. The "logistic" part of the logistic regression comes from the shape of $P(y=1 \mid x)$; inverting the logit function reveals

$$P(y=1 \mid x) = \frac{1}{1+e^{-x'\beta}} = \Lambda(x'\beta),$$

where Λ is the cdf of the *standard logistic distribution* $\text{Logistic}(0,1)$.[1]

Estimation of the logistic regression is usually done with MLE. Note that the conditional likelihood of y given x is fully specified as $\Lambda(x'\beta)^y[1-\Lambda(x'\beta)]^{1-y}$. Thus, the conditional log likelihood for one observation (x,y) is given by[2]

$$\ell_\beta(y \mid x) = y \log \Lambda(x'\beta) + (1-y)\log(1-\Lambda(x'\beta)),$$

which yields the score and Hessian

$$\dot{\ell}_\beta(y \mid x) = x[y - \Lambda(x'\beta)], \qquad \ddot{\ell}_\beta(y \mid x) = -\Lambda(x'\beta)[1-\Lambda(x'\beta)]xx'.$$

The parameter β can be estimated by maximizing the sample conditional log likelihood,

$$\hat{\beta} = \arg\max_{b \in \mathbb{R}^k} \sum_{i=1}^n \ell_b(y_i \mid x_i).$$

While we cannot write down a closed-form solution to this, we know from Theorem 6.3 that, under correct specification,

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow N(0, I_\beta^{-1}),$$

where I_β is the Fisher information matrix $I_\beta = \mathbb{E}[\dot{\ell}_\beta(y_i,x_i)\dot{\ell}_\beta(y_i,x_i)']$. In many applications in social science, however, correct specification is not a sensible assumption. Therefore, the use of the "robust" variance $\mathbb{E}[\ddot{\ell}_\beta(y_i,x_i)]^{-1} I_\beta \mathbb{E}[\ddot{\ell}_\beta(y_i,x_i)]^{-1}$, where

$$\mathbb{E}[\ddot{\ell}_\beta(y_i,x_i)] = -\mathbb{E}[\Lambda(x'\beta)[1-\Lambda(x'\beta)]xx'],$$
$$I_\beta = \mathbb{E}[\text{Var}(y \mid x)xx'] + \mathbb{E}[[P(y \mid x) - \Lambda(x'\beta)]^2 xx'],$$

is recommended (Section 6.4). In Stata, this corresponds to specifying the robust variance option: `logit y x, vce(r)`.

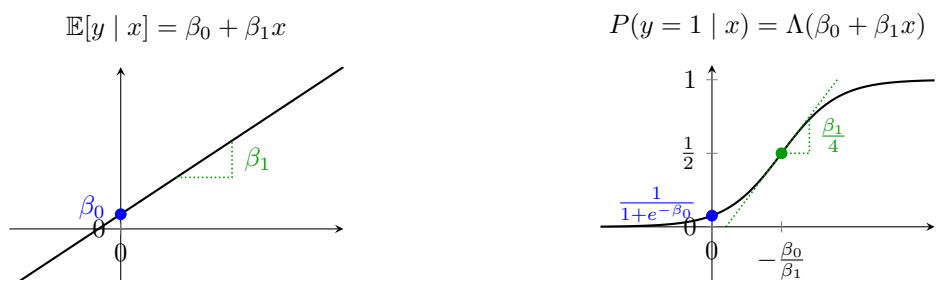
Exercise 8.1. Verify that the asymptotic variance formula in Theorem 6.6 reduces to the above expression.

[1] Note that the variance of the "standard" logistic distribution is $\pi^2/3$, not 1.

[2] Usually, the marginal density of x is assumed to not depend on the parameter, so it is irrelevant to maximization. In that case, it suffices to maximize the conditional likelihood (Remark 6.2).

TABLE 8.1. Bankruptcy prediction with financial statements [Ohl80, Table 4].

	Variable									
	SIZE	TLTA	WCTA	CLCA	NITA	FUTL	INTWO	OENEG	CHIN	Const
$\hat{\beta}$	−0.41*	6.03*	−1.43	0.08	−2.37	−1.83*	0.29	−1.72*	−0.52*	−1.32
SE	(0.11)	(0.91)	(0.76)	(0.10)	(1.28)	(0.78)	(0.35)	(0.70)	(0.24)	(1.36)



(A) Linear regression approximates the conditional expectation by a linear function.

(B) Logistic regression approximates the conditional probability (which is equal to the conditional expectation for a binary variable) by a logistic function.

FIGURE 8.1. Linear and logistic regressions.

EXAMPLE 8.1 (Prediction of bankruptcy). Accurately predicting the probability of a firm's bankruptcy is an important component of a loan approval decision for banks and investors. [Ohl80] examines the predictive power of various ratios in the financial statements. He considers nine quantities that might be indicative of default, including *SIZE* = log(total assets/GNP price-level index), *TLTA* = Total liabilities divided by total assets, and *OENEG* = $\mathbb{1}\{$total liabilities exceed total assets$\}$. The logistic regression of the indicator of bankruptcy within one year on the nine regressors and an intercept finds that five coefficients are marginally statistically significant.

8.2. Analogy and Contrast to Linear Regression

Figure 8.1 compares the logistic regression with the linear regression. Both approximates the conditional expectation of y given x, $\mathbb{E}[y \mid x]$ by a prespecified functional form. When y is binary, then the conditional expectation is equal to the conditional probability since

$$\mathbb{E}[y \mid x] = \mathbb{E}[\mathbb{1}\{y = 1\} \mid x] = P(y = 1 \mid x).$$

Since the probability is bounded by 0 and 1, the logistic regression fits a bounded function. The coefficient β_1 moves the slope of both curves and β_0 shifts the intercept. However, the slope of the fitted function changes depending on x_i, so the naive interpretation of β_1 as the overall average change of y in response to x does not hold.

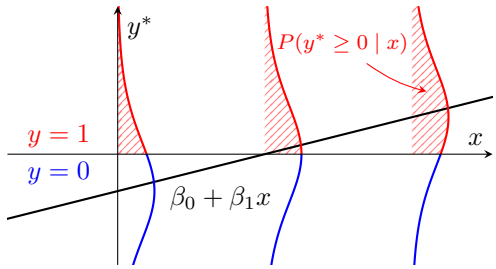


FIGURE 8.2. Latent outcome variable interpretation. The latent outcome y^* follows a logistic distribution centered at $\beta_0 + \beta_1 x$, and thus $P(y^* \geq 0 \mid x)$ is given by a logistic distribution function.

For each x, we need to calculate the slope of the function to find out the partial effect of x on $P(y = 1 \mid x)$.

Just like linear regression, we can add more than one regressor as well as their higher-order polynomial terms to make the functional form more flexible. Therefore, the logistic regression by no means forces a logistic relation between y and x; rather, it only uses a logistic relation between y and β. It may feel like logistic regression imposes more assumptions on the model as it can be estimated with MLE, but this is primarily because it exploits the binary aspect of the dependent variable (Section 8.6). Logistic regression shares the same flexibility as linear regression in designing the regressors.

8.3. Interpretations of the Logistic Regression Model

There are several data-generating processes that give rise to logistic regression.

8.3.1. Latent outcome variable. Suppose that there is a latent continuous outcome y_i^* for which the following equation holds

$$y_i^* = x_i'\beta + \varepsilon_i^*, \qquad \varepsilon_i^* \mid x_i \sim \text{Logistic}(0, 1).$$

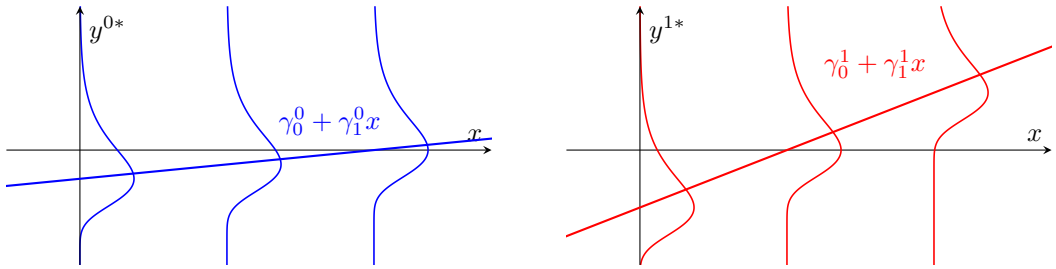
What we observe instead of y_i^* is the indicator of its positivity, that is, $y_i = \mathbb{1}\{y_i^* \geq 0\}$ (Figure 8.2). In this case, the conditional distribution of y_i is given by

$$P(y_i = 1 \mid x_i) = \Lambda(x_i'\beta).$$

In economics, for example, when the health status of an individual is only measured as *healthy* or *unhealthy*, we can think that there is an underlying continuous health status y_i^* that is unobservable to us, and an individual responds as *healthy* only when her health status is above 0. This imposes an assumption that the underlying health status is distributed as a logistic distribution conditional on observable characteristics x_i.

8.3.2. Two-way latent outcome variable. Suppose that there are two latent continuous outcome variables y_i^{0*} and y_i^{1*} such that

$$y_i^{j*} = x_i'\gamma^j + \varepsilon_i^{j*}$$



(A) $y^{0*} = x'\gamma^0 + \varepsilon^{0*}$. (B) $y^{1*} = x'\gamma^1 + \varepsilon^{1*}$.

FIGURE 8.3. Two-way latent outcome variable interpretation. The terms ε^{j*} follow a maximum Gumbel distribution. Then, the probability that $y^{1*} > y^{0*}$ is given by a logistic distribution function.

where ε_i^{0*} and ε_i^{1*} are mutually independent conditional on x_i and have the same standard type I extreme value distribution conditional on x_i. A standard type I extreme value distribution has a pdf of either $p_{\varepsilon^*}(z) = e^{-z}e^{-e^{-z}}$ (maximum Gumbel) or $p_{\varepsilon^*}(z) = e^{z}e^{-e^{z}}$ (minimum Gumbel).[3] Instead of y_i^{0*} and y_i^{1*}, we observe the indicator of their maximum, $y_i = \mathbb{1}\{y_i^{1*} > y_i^{0*}\}$ (Figure 8.3). In this case, the conditional distribution of y_i given x_i turns out to be

$$P(y_i = 1 \mid x_i) = \Lambda(x_i'(\gamma^1 - \gamma^0)).$$

Note that the two latent error terms are *not* distributed as a logistic distribution. They follow an extreme value distribution, but their difference gives rise to a logistic regression. Also, each γ is not identified; rather, their difference $\beta = \gamma^1 - \gamma^0$ is.

When we think of a consumer's decision to purchase a good, we can think that the latent utilities of purchasing and not purchasing follow an extreme value distribution conditional on observable characteristics x_i. In this case, the observed purchasing behavior is exactly modeled as the logistic regression.

By allowing more than two latent variables to choose the maximum from, we can naturally generalize this to multinomial logistic models (Section 8.7), which is often used in multiple choice models in economics.

8.3.3. Classification of two normals. Suppose that there are samples from two normal distributions with equal variance, $x_i^{1*} \sim N(\mu_1, \Sigma) =: N_1$ and $x_i^{0*} \sim N(\mu_0, \Sigma) =: N_0$. They are both observed but not labeled. Then, given a value x_i, the best guess about which distribution it came from takes the form of a logistic regression (Figure 8.4). In particular, let λ be the proportion of the sample from N_1. Define $y_i = 1$ if x_i is from N_1 and $y_i = 0$ if x_i is from N_0. The optimal classifier of x_i

[3]The variance of a standard type I extreme value distribution is $\pi^2/6$, not 1.

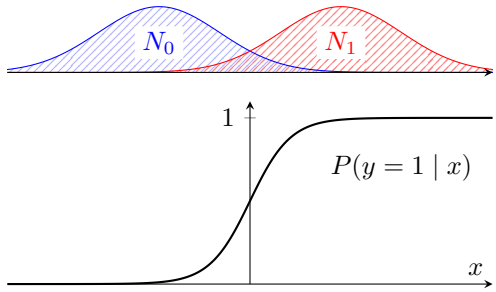


FIGURE 8.4. Classification of two normals interpretation. The sample consists of the same numbers of observations from N_0 and N_1 ($\lambda = 1/2$). Given a value x, the probability that it is drawn from N_1 is expressed as a logistic distribution function.

given y_i is equal to the conditional distribution of y_i given x_i, that is,

$$\begin{aligned} P(y = 1 \mid x) &= \frac{\lambda p_{x|y=1}(x)}{(1-\lambda)p_{x|y=0}(x) + \lambda p_{x|y=1}(x)} \\ &= \frac{1}{1 + \frac{1-\lambda}{\lambda}\exp(-x'\Sigma^{-1}(\mu_1 - \mu_0) - \frac{1}{2}(\mu_0'\Sigma^{-1}\mu_0 - \mu_1'\Sigma^{-1}\mu_1))} = \Lambda(z'\beta), \end{aligned}$$

where

$$z = \begin{bmatrix} 1 \\ x \end{bmatrix}, \qquad \beta = \begin{bmatrix} \log\frac{\lambda}{1-\lambda} + \frac{1}{2}(\mu_0'\Sigma^{-1}\mu_0 - \mu_1'\Sigma^{-1}\mu_1) \\ \Sigma^{-1}(\mu_1 - \mu_0) \end{bmatrix}.$$

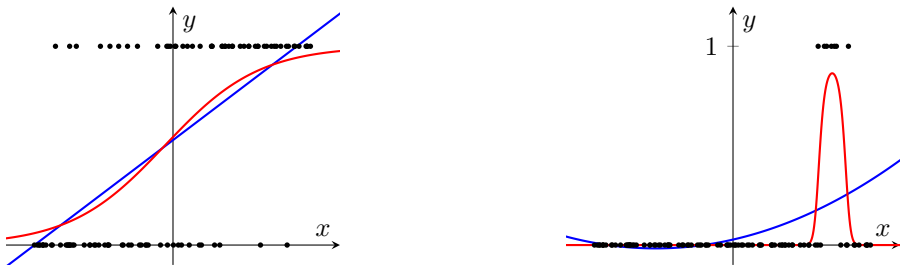
If the two normal distributions have possibly different variances, the optimal classifier takes the form of a logistic regression with regressors 1, x, and $\mathrm{vech}(xx')$. This interpretation also generalizes to multinomial logistic models by considering more than two distributions to draw the sample from (Section 8.7).

EXERCISE 8.2. Derive the explicit expression of $P(y = 1 \mid x)$ for the classification of two normal distributions with distinct nonsingular variances as well as distinct means.

Remark 8.1. More generally, the optimal classifier for arbitrary exponential family distributions is given as the (possibly nonlinear) logistic regression with the regressors being the sufficient statistics of the candidate distributions.

Remark 8.2. Technically speaking, the marginal distribution of x depends on β in this model, so maximizing the conditional likelihood of y given x is not efficient in the literal interpretation of this setup. See [**HLS13**, Section 1.5].

8.3.4. Neural network classifier. As in Section 8.3.3, a classifier is a function that takes x_i and computes the probability of $y_i = 1$ conditional on x_i. A neural network classifier gives the classifier function as the nested compositions of single-index functions in the following sense. Given an input vector x, consider the transformed variable $z_1 = \sigma(x'w_1)$ with some weight vector w_1 and a univariate function σ called the *activation function*. We can create more transformations $z_2, \ldots, z_k$ for different



(A) For monotonic and balanced data, the linear and logistic regressions give similar predictions. The blue line is the linear regression $y = \beta_0 + \beta_1 x + \varepsilon$, and the red line the logistic regression $P(y = 1 \mid x) = \Lambda(\beta_0 + \beta_1 x)$.

(B) For nonmonotonic or imbalanced data, the linear and logistic regressions give different predictions. The blue line is the linear regression $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, and the red line the logistic regression $P(y = 1 \mid x) = \Lambda(\beta_0 + \beta_1 x + \beta_2 x^2)$.

FIGURE 8.5. Difference of the LPM and logistic regression.

weights $w_2, \ldots, w_k$ but usually with the same activation function. These variables $z_1, \ldots, z_k$ constitute the output of the first layer called a hidden layer. Then, having z as the new input, we can consider the further transformations of the same type. After some iterations, we then compute the final output by $\Lambda(\tilde{z}'w)$ where Λ is the cdf of the logistic distribution and w a weight vector. Thus, a neural network consists of the initial input layer, several hidden layers, and the final output. The cdf of the logistic distribution is called the sigmoid function in this context, and the activation functions for the hidden layers may also be taken as the sigmoid function.

Logistic regression can then be interpreted as the neural network classifier with no hidden layer. That is, given the input vector x, we directly compute the final output as $\Lambda(x'\beta)$.

8.4. Relation to the Linear Probability Model

Even when the dependent variable is binary, we can still run the linear regression of y on x since the linear regression is always well defined (Section 7.1.2). This is called the *linear probability model (LPM)*. That is, even when the linear regression line may not make sense as the conditional probability of $y = 1$ given x, it still does as the best linear approximation to it in the sense of the mean squared error. In many situations the linear probability model is clearly misspecified (e.g., when the support of x is unbounded), but for that matter, any model is misspecified to some extent. Linear regression may give us predictions outside of 0 and 1, but in turn enjoys unequivocal interpretability, ease of comparison across studies, and familiarity to many practitioners. For this reason, some authors prefer to use the linear probability model even when the outcome variable is binary. In fact, it is a common observation in practice that the partial effect estimates using the logistic regression and the linear regression coincide to a good degree. This is especially so when the numbers of $y_i = 1$ and $y_i = 0$ are of a comparable order (Figure 8.5A). See also [**AP09**, Section 3.4.2] for this matter.

However, when there is severe imbalance in the numbers of observations of different categories (for example, customers' purchase indicators of products on Amazon consist mostly of zeros, and so do the default indicators of credit cards), the linear probability model can fail miserably. Also, when the conditional probability function is not monotonic, linear and logistic regressions tend to yield very different results (Figure 8.5B). In these cases, models that take into account the bounded property of y and interpolate the data with a nice smooth curve would prove much more useful.

Also, when the dependent variable is multinomial, the linear regression is considered a terrible practice. Linear regression forces us to place the categories on a real line, which simply does not make sense in many cases. For example, consider a mover company that develops a web system that provides a plan recommendation for customers in response to the inputs such as locations and how much furniture they have to move. Different plans may be tailored for different types of customers in various dimensions, and cannot simply be assigned a number that changes linearly with the customers' diverse characteristics. In such cases, logistic regression (or probably a more sophisticated machine learning method) would be a better option.

8.5. Relation to the Probit Model

In addition to adding regressors, another way to make the method more flexible is to use a function other than Λ. In general, the function that connects the conditional probability of y to the index $x'\beta$ is called the *link function*, and it is for us to choose. For example, we can set the link function to be the standard normal cdf,

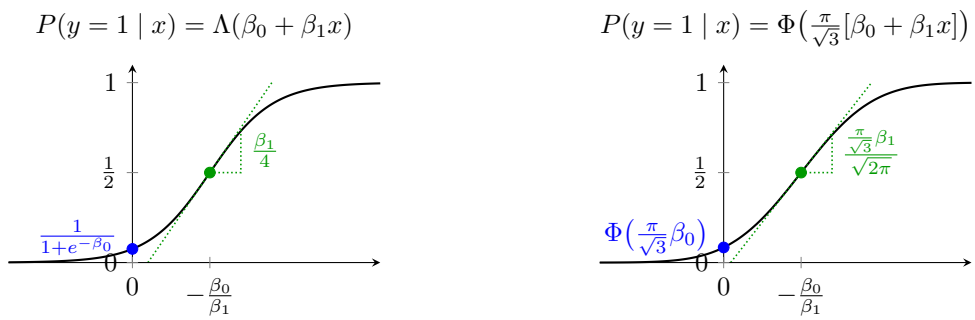
$$P(y = 1 \mid x) = \Phi(x'\beta),$$

in which case the model is called the *probit* model. This has a similar interpretation as Section 8.3.1 when $\varepsilon_i^* \mid x_i$ follows a standard normal distribution.

When the binary outcome is concerned, the probit model seems to win the popularity vote in economics over the logistic regression (also known as the *logit model* there), probably because the normal latent outcome interpretation is easy to conceive. In statistics and computer science, however, logistic regression has acquired the position of the canonical (most basic) classification method. In fact, logistic regression has many advantages over the probit model. First, it admits a closed-form likelihood as the cdf of a logistic distribution has a closed-form expression. Second, it has various interpretations that speak to diverse practitioners (Section 8.3), and for that matter, it also relates to the normal distribution in a neat way (Section 8.3.3). Third, it naturally extends to multinomial responses in an interpretable way (Section 8.7). For this last reason, when it comes to multiple choice models, multinomial logistic regression is the go-to method in economics as well.[4]

Remark 8.3. It is known that there is almost a one-to-one relationship between the coefficients of the logit and the probit models; the logit β is roughly equal to 1.6 to

[4]There is a multinomial extension of the probit model [**BAL85**, Section 5.7], but it is computationally intensive or intractable.



(A) Logit model. (B) Probit model.

FIGURE 8.6. Logit and probit models. Multiplying $\pi/\sqrt{3}$ to the coefficients makes the probit model very close to the logit. This corresponds to equating the variance of the latent error terms.

1.7 times the probit β [**Gre18**, Chapter 17]. This is mainly due to the difference in the variances of the standard logistic and standard normal distributions (Figure 8.6).

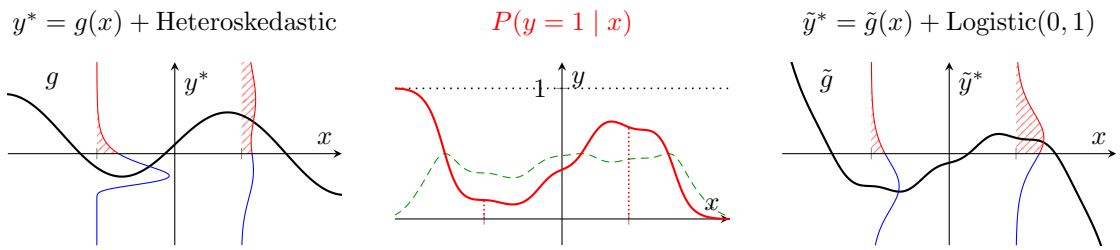
8.6. Is There "Heteroskedasticity?"

In the latent outcome interpretation (Section 8.3.1), the distribution of the "error term" ε_i^*—let alone its variance—does not depend on x_i. Does this mean that the logistic regression is "confined" to a homoskedastic framework? In fact, this is a false analogy with the linear regression when it comes to heteroskedasticity.

Suppose that there is a "true" latent outcome y^* that is heteroskedastic, i.e., $y^* = g(x) + \varepsilon^*$ and $\varepsilon^* \mid x$ follows a varying distribution that depends on x (Figure 8.7A). This induces a conditional probability function on the observed outcome y through $P(y = 1 \mid x) = P(y^* \geq 0 \mid x)$ (Figure 8.7B). This in turn gives rise to an alternative latent outcome $\tilde{y}^*$ such that $\tilde{y}^* = \tilde{g}(x) + \tilde{\varepsilon}^*$ for $\tilde{\varepsilon}^* \mid x \sim \text{Logistic}(0, 1)$ and $P(\tilde{y}^* \geq 0 \mid x) = P(y = 1 \mid x)$ (Figure 8.7C). Therefore, as long as the log-odds ratio $x'\beta$ is modeled flexibly enough to approximate $\tilde{g}$ well, there is no loss of generality in modeling the logistic latent error. In a nutshell, the latent outcome is not doomed to be equivariant, but to our advantage can be deemed so. It is only when we try to directly interpret β in the context of the structural model (such as welfare analysis) that the distributional assumption becomes material and entails loss of generality.[5]

Additionally, since y is a binary variable, the error term of the observed outcome $\varepsilon = y - P(y = 1 \mid x)$ is a demeaned Bernoulli variable, so there is no "flexibility" in the distribution of ε and heteroskedasticity is a necessity. In particular, the variance of a Bernoulli random variable with probability p is given by $p(1 - p)$, so the conditional variance of y given x is a deterministic function of the conditional mean, namely, $P(y = 1 \mid x)[1 - P(y = 1 \mid x)]$ (Figure 8.7B). The logistic regression does not need to

[5]Since utility is ordinal, the distributional assumption is innocuous at the individual level even with a structural interpretation [**BAL85**, Section 4.1]. However, if we aggregate utilities over different individuals, it imposes an assumption on their comparability. This interpretational issue is a chronic problem of structural models, and the LPM is not free of it either.



(A) True latent outcome y^* with heteroskedastic error.

(B) Conditional probability of observed outcome y.

(C) Deemed latent outcome $\tilde{y}^*$ with logistic error.

FIGURE 8.7. "True" latent outcome y^* may be heteroskedastic, but there exists a "deemed" latent outcome $\tilde{y}^*$ with a standard logistic error that induces the same conditional probability function on the observed outcome y, i.e., $P(y = 1 \mid x) = P(y^* \geq 0 \mid x) = P(\tilde{y}^* \geq 0 \mid x)$. In this sense, the distributional assumption on the latent error is innocuous. Conditional heteroskedasticity of y is then determined solely by the conditional probability function (green dashed line plots $\sqrt{\text{Var}(y \mid x)}$).

explicitly take this heteroskedasticity into account since the conditional likelihood of y accounts for the entire probabilistic structure of the Bernoulli process, not only the heteroskedasticity. Of course, the use of the robust standard error is recommended (Section 8.1), but this is to robustify the inference against *misspecification*, not against heteroskedasticity.

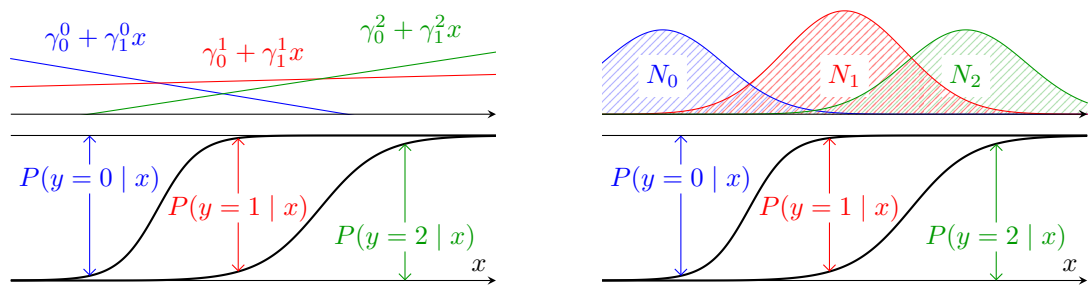
Meanwhile, if we go with the linear probability model, the use of a heteroskedasticity-robust standard error is a must, regardless of the presence of misspecification. Linear regression remains agnostic on the specificity of the binary outcome and allows much flexibility in the error distribution—in this case it just happens to be vacuous—so we need to explicitly account for the heteroskedasticity in the standard error calculation.

8.7. Multinomial Logistic Regression

When y takes on more than two categories, it is important to distinguish two types of categories. The *nominal* categories are not associated in a particular order, e.g., States, colors, and products sold at a supermarket. The *ordinal* categories are ordered in a meaningful way, e.g., ratings on Yelp, course grades, and clothing sizes.[6] The extension of logistic regression for nominal y is called the *multinomial logistic regression*,[7] and the extension for ordinal y is called the *ordinal* (or *ordered*) *logistic regression*. This section gives an introduction to the simpler one, multinomial logistic regression. For ordinal logistic regression, see [**HLS13**, Chapter 8] or [**Woo10**, Chapter IV]. See also [**BAL85**, Chapter 10] for other kinds of logistic regression models.

[6]Of course, there are cases where there is only a meaningful *partial* order, as in job titles.

[7]Not to be confused with *multivariate* logistic regression, which simply is a binomial logistic regression with multiple regressors.



(A) Three-way latent outcome variables interpretation.

(B) Classification of three normals interpretation.

FIGURE 8.8. Multinomial logistic model.

Suppose that y takes on m categories labeled $\{0, 1, \ldots, m-1\}$. Multinomial logistic regression models the relation of y and x by

$$\log \frac{P(y = \ell \mid x)}{P(y = 0 \mid x)} = x'\beta^\ell$$

for each $\ell = 1, \ldots, m-1$. Note that there is a separate set of coefficients β^ℓ for each category except for 0, so the total number of coefficients is $k(m-1)$. Solving the system of log-odds ratios yields

$$P(y = 0 \mid x) = \frac{1}{1 + \sum_{l=1}^{m-1} e^{-x'\beta^l}}, \qquad P(y = \ell \mid x) = \frac{e^{-x'\beta^\ell}}{1 + \sum_{l=1}^{m-1} e^{-x'\beta^l}}.$$

Again, β can be estimated by maximizing the conditional log likelihood.

The generalization of the two-way latent outcome variables interpretation in Section 8.3.2 is a handy way to interpret this model (Figure 8.8A). Suppose that there are m latent outcomes $y_i^{0*}, \ldots, y_i^{(m-1)*}$, each of which is generated as

$$y_i^{\ell*} = x_i'\gamma^\ell + \varepsilon_i^{\ell*}, \qquad \varepsilon_i^{\ell*} \mid x_i \sim \mathrm{EV}_1(0, 1).$$

We observe the index of the maximum of $y_i^{0*}, \ldots, y_i^{(m-1)*}$, that is, $y_i = \arg\max_\ell y_i^{\ell*}$. With this, the conditional probability of y_i given x_i becomes

$$P(y_i = \ell \mid x_i) = \frac{e^{-x'(\gamma^\ell - \gamma^0)}}{1 + \sum_{l=1}^{m-1} e^{-x'(\gamma^l - \gamma^0)}},$$

which coincides with the above modeling through $\beta^\ell = \gamma^\ell - \gamma^0$.

This interpretation nicely fits the multiple choice models in economics. When a consumer has a multiple set of choices, the consumer chooses the one that maximizes her utility. If the utility from each choice is given as a random variable distributed as a type I extreme value distribution centered at some function of the observed characteristics, then her utility function can be estimated by the multinomial logistic regression, which, for example, enables welfare analysis [**WF94**]. Unlike the binary outcome case, however, the assumption of independent logistic errors is not without loss of generality; the logistic multiple choice model is known to satisfy the axiom of "independence from irrelevant alternatives (IIA)," which simplifies the analysis but

sometimes leads to paradoxical behaviors [**BAL85**, Section 5.3]. [**MM15**] justify the use of multinomial logistic regression for multiple choice models based on rational inattention. In the context of revenue management, the logit choice model admits an efficient algorithm to optimize assortment [**TvR04**, Proposition 6].

We can also interpret the multinomial logit model through the classification of multiple normal distributions in the same spirit as Section 8.3.3 (Figure 8.8B). Let $p_0, \ldots, p_{m-1}$ be m pdfs from which an observation is drawn randomly with respective probabilities $\lambda_0, \ldots, \lambda_{m-1}$. Let $y = \ell$ if x comes from the pdf p_ℓ. The optimal classifier for this model is then

$$P(y = \ell \mid x) = \frac{\lambda_\ell p_\ell(x)}{\lambda_0 p_0(x) + \cdots + \lambda_{m-1} p_{m-1}(x)}.$$

If we let each p_ℓ be a normal pdf with equal variance, we obtain the multinomial logistic regression where β is determined by the means and variance of the normals. It is straightforward to generalize it to arbitrary exponential family distributions.

8.8. Nonparametric Classification

Just like regression, classification can be considered in a nonparametric way. As explained in Section 8.6, we can make the index $x'\beta$ nonparametric to accommodate all possible conditional probability functions. For example, we can add higher-order polynomials or other transformations [**HIR03**]. More terms can be added as more observations are available, so that in the limit we can "saturate" the model just as in linear regression. Typical neural network classifiers can also be understood in this framework. Usually, the output layer is given with the sigmoid activation function, which corresponds to the link function arising from the logistic regression. Then the function specified as the input to the last output layer (their activation function need not be sigmoid) is the nonparametrically specified index function.

Another direction is to make the link function nonparametric while assuming that the index part is known or at least parametrically specified. This corresponds to allowing an arbitrary distribution for the latent error term ε^* in the one-way latent outcome interpretation (Section 8.3.1). [**KS93**] consider a semiparametrically efficient estimator for this model. Note that, unlike making the index nonparametric, this "semiparametrization" cannot accommodate all possible conditional probability functions. However, when we have a strong reason to believe that the index specification is known up to a finite-dimensional parameter, this model can recover the "true" latent outcome model that can then be used for further structural analyses.

We can think of classification in the framework other than MLE. In the machine learning literature, random forests and support-vector machine are popularly used for nonparametric classification problems. Also, the Wasserstein loss is frequently used to train a neural network classifier for generative adversarial networks (GAN) to avoid problems arising from the disjoint supports of high-dimensional datasets.

CHAPTER 9

Principles of Causal Inference

Random number generation is too important to be left to chance.

Robert Coveyou, 1969

9.1. Correlation Does Not Imply Causation?

An economist may want to know how provision of health care improves the mental health of an individual. A corporate finance researcher may want to examine how the capital structure influences firm investments. A student of accounting may want to inquire how the International Financial Reporting Standards affect liquidity. A marketer may want to quantify how much a dollar spent on an ad increases sales. A government agency may want to investigate how the selective advertising by Facebook alters voters' behaviors.

These questions are *causal* in nature; they all concern how one thing leads to another. This is more than a correlation. Simply finding that individuals with health care tend to have better mental health does not mean that health care improved their mental health; it may just be that those with better mental health tend to choose to get health care. This problem is encapsulated in the famous saying: "Correlation does not imply causation." In social science, "everything correlates to some extent with everything else" [Mee90].

However, as statistics being a practical application of probability theory, all that statistics can handle is correlation. There is no Stata command that says "`cause y x`" and gets you a causal effect. Does this mean that statistics is helpless and we

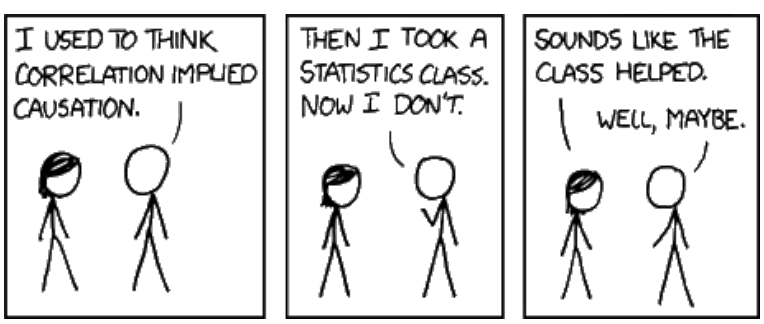


Figure 9.1. `https://xkcd.com/552`

never get to causality? The answer is fortunately no—under certain conditions, a certain kind of correlation can be interpreted as causation.[1] The branch of statistics that studies it is known as *causal inference* and *program evaluation*.

In general, there are three approaches to scientific inquiry of causality.

(1) A *deductive* approach aims to uncover the causal chain of reactions. A neuroeconomic research on how decisions are made in our brains is primarily deductive.
(2) An *inductive* approach aims to statistically discover causality while remaining agnostic on the underlying chain of reactions. Reduced-form modeling is primarily inductive.
(3) An *abductive* approach hypothesizes the chain of reactions and examines counterfactual scenarios under the postulated hypothesis. Structural modeling is primarily abductive.[2]

In this chapter, we discuss principles of the inductive approach to causality. For further reading, I recommend [**Pea09**, **MW14**, **IR15**, **HR20**].

9.2. The Inductive Model of Causality

We first need to define causality. Let us look at an example of drug evaluation. If you are to define what it means for the drug to be effective in curing some disease, you may think of something like this: the drug contains these molecules, and when taken orally, they are absorbed into this organ, and the organ reacts such and such. While this is a totally fine way to define causality, this is not the only way to define it. Surprisingly, we can define the causal effect of the drug without any knowledge about biology or medicine.

Consider the following thought experiment. Suppose there are two parallel worlds, A and B. These worlds are very much like each other; in fact, there is one and only one difference between the two. In world A, a patient does not take this drug; in world B, the same patient does take this drug. That's it. Everything else, even the perception of the patient, is the same across the two, so if the patient "thinks" that he is taking the new drug in world B, so does he in world A. After a while, we observe the health status of this patient in both worlds. If the patient's disease is cured in B but not in A, we can attribute it to the drug, for that was the only difference that distinguished the two parallel worlds to begin with.

The advantage of this definition is that there is no need to know *how* the drug helped cure the disease. This is in fact a major advantage in social science because the specific mechanism by which one thing leads to another is hardly known for sure. The disadvantage is that there are no two parallel worlds for us to compare, and this is where statistics comes into play. While it is certainly not possible to replicate our

[1]For this, I prefer saying "Correlation does not *always* imply causation"—as we will see, some correlations do.

[2]Obviously, the given examples are an oversimplification; estimation of neural signals in neuroeconomics is inductive; imposing an exclusion restriction in reduced-form modeling is abductive; economic theory that endorses structural modeling is deductive.

world and create a situation in which "everything else" is the same, it *is* possible to replicate the situation in which the "probability distribution of everything else" is the same. Say that you generate a Bernoulli random variable on your laptop. It is by construction independent of anything else in the world; the conditional distribution of everything else given the outcome of your Bernoulli variable—be it 0 or 1—is the same as the marginal distribution of everything else (Exercise 2.15). Therefore, if you assign the drug depending on the outcome of the Bernoulli variable and do not let any other thing depend on it (e.g., you don't tell which drug was assigned to the patient or even to the doctor), you can create a situation in which the "distribution of everything else" is the same for either drug assignment.

Of course, we cannot observe both outcomes with and without the drug for one patient. However, as far as the *average* effect of the drug is concerned, this poses no problem. The average outcome of the drug for a population can be estimated by the average outcome of the drug for a *randomly chosen subset* of the population and its convergence is very fast. Combined with the similarly estimated average outcome of not taking the drug, we can estimate the average causal effect of the drug by taking their difference.

Thus, the key to causal inference is to find the variation that is independent of the causal chain in question. As discussed above, the easiest method to find such a variation is to generate one yourself. This is called the *(controlled) experiment* or the *randomized controlled trial (RCT)*. In social science, however, it is not always possible to run an experiment, for reasons such as costs or ethics. Even then, we can sometimes find an exogenous variation in observational data as if an experiment took place. Such cases are called the *quasi-experiments* or the *natural experiments*.

Finally, the word "random" in this context is often used to mean a probabilistic variation that is independent of the causality in question. This is in contrast to our use in earlier chapters as a synonym for "probabilistic."

9.3. Causal Inference with Experimental Data

Let us introduce the notation. First, we denote a subject by i. In the health care example, i refers to an individual; in the corporate finance example, it refers to a firm. Whether the subject i is assigned the treatment is denoted by X_i, so $X_i = 1$ if i receives the treatment and $X_i = 0$ if not. The outcome in the scenario where i receives the treatment is denoted by Y_{i1}, and likewise Y_{i0} if i does not receive the treatment. By all means, we observe only one of Y_{i1} and Y_{i0} for subject i; the actually observed outcome is given by $Y_i = X_i Y_{i1} + (1 - X_i) Y_{i0}$. The difference $Y_{i1} - Y_{i0}$ for a specific subject i is called the *individual treatment effect*. The mean of it $\mathbb{E}[Y_{i1} - Y_{i0}]$ is the *average treatment effect (ATE)*, which is our primary target.

9.3.1. Experiments with random treatment. When the treatment assignment X_i can be set randomly, estimation of the causal effect is straightforward. Suppose we run the following regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

This equation can be understood as two equations:

$$\begin{cases} Y_{i0} = \beta_0 + \varepsilon_i & \text{if } X_i = 0, \\ Y_{i1} = \beta_0 + \beta_1 + \varepsilon_i & \text{if } X_i = 1. \end{cases}$$

This implies that $\beta_0 = \mathbb{E}[Y_{i0}]$ and $\beta_1 = \mathbb{E}[Y_{i1}] - \mathbb{E}[Y_{i0}]$. Ergo, the OLS coefficient $\hat{\beta}_1$ estimates the ATE.

Note that including other regressors that are uncorrelated with X_i in the equation does not change our target parameter. Recall the discussion right before Theorem 7.11. First, the coefficient γ_1 in the regression

$$Y_i = \gamma_0 + \gamma_1 X_i + W_i' \gamma_2 + \varepsilon_i$$

is the same as the coefficient γ_1 in

$$Y_i - \mathbb{E}[Y_i] = \gamma_1 (X_i - \mathbb{E}[X_i]) + (W_i - \mathbb{E}[W_i])' \gamma_2 + \varepsilon_i.$$

Then, it is the same as γ_1 in

$$M_W(Y_i - \mathbb{E}[Y_i]) = \gamma_1 M_W(X_i - \mathbb{E}[X_i]) + \varepsilon_i,$$

where $M_W(z) \coloneqq z - (W_i - \mathbb{E}[W_i])' \operatorname{Var}(W_i)^{-1} \operatorname{Cov}(W_i, z)$. However, since X_i is independent of W_i, we get $\operatorname{Cov}(W_i, X_i) = 0$. So the equation reduces to

$$M_W(Y_i - \mathbb{E}[Y_i]) = \gamma_1 (X_i - \mathbb{E}[X_i]) + \varepsilon_i.$$

Therefore,

$$\gamma_1 = \frac{\mathbb{E}[(X_i - \mathbb{E}[X_i]) M_W(Y_i - \mathbb{E}[Y_i])]}{\operatorname{Var}(X_i)} = \frac{\mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_i - \mathbb{E}[Y_i])]}{\operatorname{Var}(X_i)}$$

since, again, $\operatorname{Cov}(W_i, X_i) = 0$. This matches the formula for β_1,

$$\beta_1 = \frac{\operatorname{Cov}(X_i, Y_i)}{\operatorname{Var}(X_i)}.$$

However, their estimates are not numerically the same; if W_i is a set of relevant variables for Y_i, $\hat{\gamma}_1$ may be more precise than $\hat{\beta}_1$ since W_i absorbs variation of Y_i not attributable to X_i; if W_i is irrelevant, $\hat{\gamma}_1$ may be less precise since W_i takes away some degrees of freedom.

Covariates can also be used to check the quality of randomization. If X_i is independent of everything, the conditional distribution of W_i given X_i does not depend on X_i. Therefore, we can test the quality of randomization by testing if the conditional distribution of W_i is invariant of X_i (*covariate balance*). A popular choice is to compare the mean and standard deviation of W_i across groups $X_i = 1$ and $X_i = 0$, but other characteristics (or even the distribution itself) can also be used. We can check the covariate balance right after randomization before applying the treatment. If we detect severe imbalance due to "bad" randomization, we can randomize again.

Example 9.1 (Reaching for yield). In finance, *reaching for yield* refers to the behavior that an investor faced with a low average return tends to undertake high risk. Some suspect that such a behavior has contributed to financial crisis. [**LMW19**] carry out an experiment and investigate whether such a behavior exists in MBA

TABLE 9.1. Covariate balance [LMW19, Table 1].

		Treated		Control		Difference	
		N	%	N	%	%	p-value
Gender	Male	117	58.2	129	64.8	−6.7	0.17
	Female	84	41.8	70	35.2	6.7	
Risk tolerance	High	116	57.7	107	53.8	3.9	0.55
	Medium	48	23.9	56	28.1	−4.3	
	Low	37	18.4	36	18.1	0.3	
Investment experience	More	93	46.3	85	42.7	3.6	0.47
	Limited	108	53.7	114	57.3	−3.6	
Worked in finance	Yes	84	41.8	86	43.2	−1.4	0.77
	No	117	58.2	113	56.8	1.4	
Total		201		199			

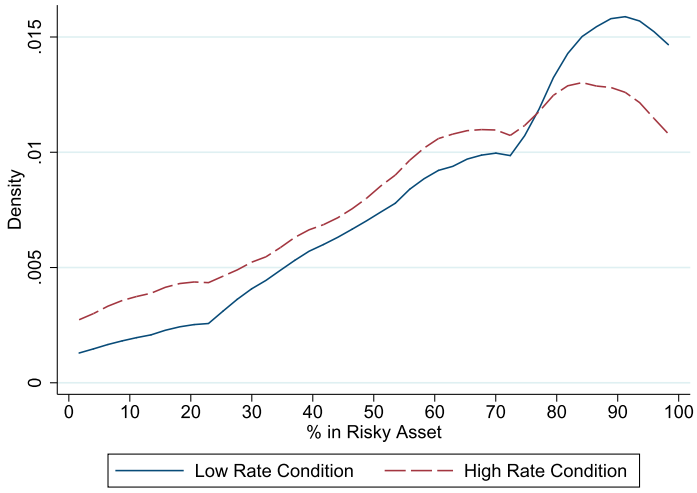


FIGURE 9.2. Distributions of allocation to risky asset [LMW19, Figure 2].

students at Harvard Business School. The MBA students are asked to allocate a hypothetical endowment of 1,000,000 Francs to the risk-free asset and the risky-asset. They are randomly selected into two groups, A and B. Group A is presented with the two assets: (1) the risk-free asset with 5% return with 0% risk and (2) the risky asset with 10% return and 18% risk. Group B: (1) the risk-free asset with 1% return with 0% risk and (2) the risky asset with 6% return and 18% risk. Thus, Group B was exposed to a lower average return compared to A but with the same risk profile. Table 9.1 checks the distribution of covariates across groups and lists p-values that test the hypotheses of no difference using the Mann–Whitney U test. Figure 9.2 presents the density estimators of the distributions of the risky asset allocation (red

TABLE 9.2. Reaching for yield experiment [LMW19, Table 2].

		% risky allocation			
		(1)	SE	(2)	SE
	Low return	8.83*	(2.82)	8.76*	(2.75)
Control W_i	Male			6.25*	(2.92)
	Risk tolerance medium			5.56	(4.09)
	Risk tolerance high			15.39*	(3.84)
	Investment experience			4.41	(3.45)
	Worked in finance			3.34	(3.34)
	Constant	66.79		55.81*	(3.91)
	N	400		400	

for the control group and blue for the treatment group). The authors then run the regression

$$RiskyAlloc_i = \beta_0 + \beta_1 LowReturn_i + \varepsilon_i,$$

where $RiskyAlloc_i$ is the percentage of allocation to the risky asset by an individual i and $LowReturn_i$ indicates whether i was in group B (Table 9.2 (1)). They find that $\hat{\beta} = 8.83$ with the standard error 2.82, yielding a significant coefficient for the reaching for yield estimate. They also estimate the regression including demographic controls such as the risk tolerance level, investment experience, and work experience in financial industry (Table 9.2 (2)). With that, reaching for yield is estimated more precisely at 8.76 with standard error 2.75. This behavior does not arise from conventional portfolio choice theory or institutional frictions, and they provide possible mechanisms of investor psychology that can explain this such as reference dependence and salience.

EXAMPLE 9.2 (Therapeutic effects of intercessory prayer). Intercessory prayer is widely believed to influence recovery from illness. [BDS+06] examine whether the prayer itself or the knowledge thereof influences recovery. They took as the subjects 1,802 patients who plan to undergo the coronary artery bypass graft (CABG) surgery, which is susceptible to about 50% chance of complications and 5% of death. They were randomly assigned to three groups: (1) those who were informed that they would receive the prayer, and received the prayer, (2) those who were informed that they may or may not receive the prayer, and received the prayer, and (3) those who were informed that they may or may not receive the prayer, and did not receive the prayer. Thus, the difference between (1) and (2) reveals the causal effect of acknowledging the prayer, and the difference between (2) and (3) the causal effect of receiving the prayer (Figure 9.3). The outcome measure is whether the patient suffers complications within 30 days of the surgery. They find that the complication rates for the three groups were (1) 58.6%, (2) 52.2%, and (3) 50.9%. The estimated ATE for acknowledging the prayer on the complication rate is 6.4% with the standard

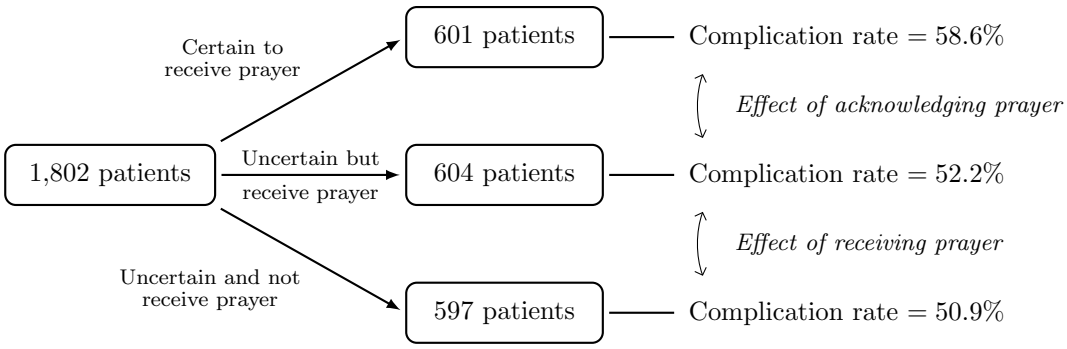


FIGURE 9.3. Randomization into three groups and two treatment effects considered in [BDS+06].

error of 2.86% (significant with size 5%), and the ATE for receiving the prayer on the complication rate is 1.2% with the standard error of 2.89% (insignificant).

EXAMPLE 9.3 (Network effect). Humans are not good at making rational decisions in a long horizon. For example, we may not think seriously enough about the election of retirement benefits at an early stage of our career. [DS03] investigate whether short-term monetary incentives can increase attendance to a fair and affect the choice of retirement benefits. They designed a clever experiment to separately identify the effects of incentives and of networks. For example, if we randomly distribute a small reward to attend a fair to employees, some of the treated individuals may spread the word. It may encourage those in the control group to attend the fair with them, or it may discourage those to do so had they found out that they were not eligible for the reward. To distinguish such effects from the effect of the rewards, they carried out the experiment in two steps. First, they randomly classified 330 departments in a university into two groups: 220 "treated" departments and 110 "control" departments. Among the 4,168 employees in the treated departments, they randomly sent out an invitation letter promising a \$20 reward to 2,039 employees for attending the fair. The remaining 2,129 employees in the treated departments and the 2,043 employees in the control departments did not receive the letter. Thus, the individuals are classified into three groups: (1) those who are in the treated departments and received the incentive, (2) those who are in the treated departments and did not receive the incentive, and (3) those who are in the control departments and did not receive the incentive (Figure 9.4). The difference between (1) and (2) identifies the effect of receiving the incentive, and the difference between (2) and (3) the *network effect* or the *peer effect*. The regression model is, therefore,

$$Y_i = \beta_0 + \beta_1 \mathit{TreatedDept}_i + \beta_2 \mathit{Treated}_i + \varepsilon_i,$$

where $\mathit{TreatedDept}_i = 1$ if i is in the treated department, and $\mathit{Treated}_i = 1$ if i received the letter. Various outcomes are examined, including the fair attendance and the Tax Deferred Account (TDA) enrollment within 4.5 months after the fair (Table 9.3). For fair attendance, $\hat{\beta}_1$ and $\hat{\beta}_2$ were both statistically significant at 0.102 and 0.129; this is to say that being in the treated department increases the chance of

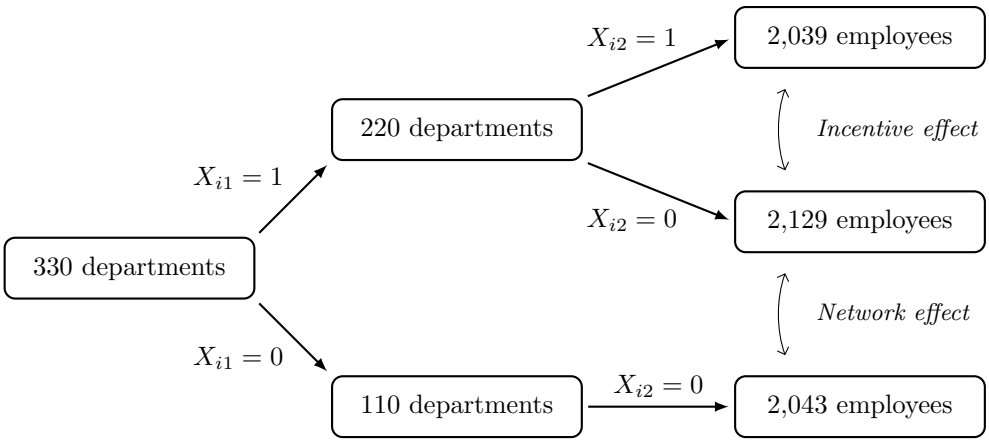


FIGURE 9.4. Two-step randomization in [DS03]. $X_{i1} = 1$ indicates that some employees in i's department received the reward email; $X_{i2} = 1$ that i received the reward email.

TABLE 9.3. Role of information and peer influence [DS03, Table II].

	Outcome			
	Fair attendance		TDA enrollment after 4.5 months	
Treated department	0.102*	(0.014)	0.013*	(0.005)
Treated	0.129*	(0.023)	−0.007	(0.006)
N	6,144		5,587	

attending the fair by 10.2% compared to the employees in the control departments, and receiving the reward further increases the chance by 12.9%. In relative terms, the fair attendance was five times higher for (1) and three times higher for (2) than for (3). For TDA enrollment, $\hat{\beta}_1$ was significantly positive at 0.013 while $\hat{\beta}_2$ was not significant. Therefore, the word of mouth was effective in promoting TDA enrollment but the monetary incentive was not. The authors then hypothesize what might have been happening: (1) those who attended the fair spread information; (2) those who attended the fair for incentives were different from those who attended the fair for their colleagues; (3) receiving the reward reduced motivation.

9.3.2. Experiments with random eligibility. In many applications, it is not possible to randomly assign the treatment. If an economist wants to evaluate a job training program, it may be possible to randomly send an invitation to unemployed individuals, but it is not possible to force someone to take the program if he is not willing to. If a marketer wants to evaluate a new subscription plan, she will be able to promote the plan to randomly selected customers but not force them to purchase it. Thus, a common situation in social science is that *eligibility* to receive the treatment

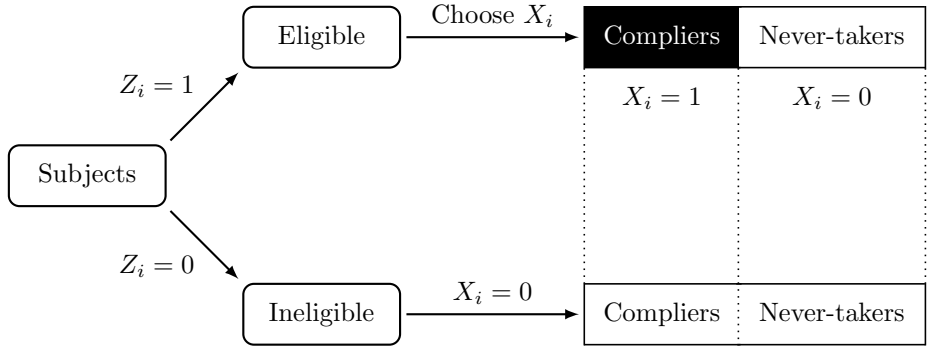


FIGURE 9.5. Experiments with random eligibility. Since eligibility is randomly assigned, the distribution of potential compliers and never-takers among ineligible subjects is the same as the distribution of observed compliers and never-takers among eligible subjects.

can be allotted randomly, while the decision to participate is left to the subject. Even in that case, it is possible to estimate the causal effect of the treatment, albeit limited to some subpopulation.

Let us introduce a variable Z_i that indicates whether subject i is eligible to participate in the treatment. In our setting, Z_i is randomly assigned either 0 or 1. If $Z_i = 1$, subject i then chooses whether to participate $X_i = 1$ or not $X_i = 0$. If $Z_i = 0$, subject i has no option to participate, so $X_i = 0$ (Figure 9.5).

First, note that it is straightforward to estimate the causal effect of *eligibility*. If we run the regression

$$Y_i = \alpha_0 + \alpha_1 Z_i + \varepsilon_i,$$

the coefficient $\alpha_1 = \mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]$ estimates the average causal effect of receiving the invitation to the treatment on the outcome, regardless of whether they end up taking it. This is called the *intent-to-treat (ITT)* effect.

To understand how we can get to the causal effect of the treatment, it is helpful to decompose ITT into two groups of subjects. Let us call the subjects who would opt in to the program had they been eligible the *compliers*, and those who would opt out the *never-takers*. Then, $\mathbb{E}[Y_i \mid Z_i = 1]$ can be decomposed as

$$\begin{aligned}\mathbb{E}[Y_i \mid Z_i = 1] &= P(\text{complier})\mathbb{E}[Y_i \mid Z_i = 1, \text{complier}] \\ &\qquad + P(\text{never-taker})\mathbb{E}[Y_i \mid Z_i = 1, \text{never-taker}] \\ &= P(\text{complier})\mathbb{E}[Y_{i1} \mid \text{complier}] + P(\text{never-taker})\mathbb{E}[Y_{i0} \mid \text{never-taker}].\end{aligned}$$

Here, I used the observation that $Y_i = Y_{i1}$ for compliers with $Z_i = 1$, $Y_i = Y_{i0}$ for never-takers with $Z_i = 1$, and Z_i was randomly assigned (so can be removed from conditioning). For simplicity, let $\pi := P(i \text{ is a complier})$. Then, ITT is the difference of the following two quantities.

$$\begin{aligned}\mathbb{E}[Y_i \mid Z_i = 1] &= \pi\mathbb{E}[Y_{i1} \mid \text{complier}] + (1-\pi)\mathbb{E}[Y_{i0} \mid \text{never-taker}], \\ \mathbb{E}[Y_i \mid Z_i = 0] &= \pi\mathbb{E}[Y_{i0} \mid \text{complier}] + (1-\pi)\mathbb{E}[Y_{i0} \mid \text{never-taker}].\end{aligned}$$

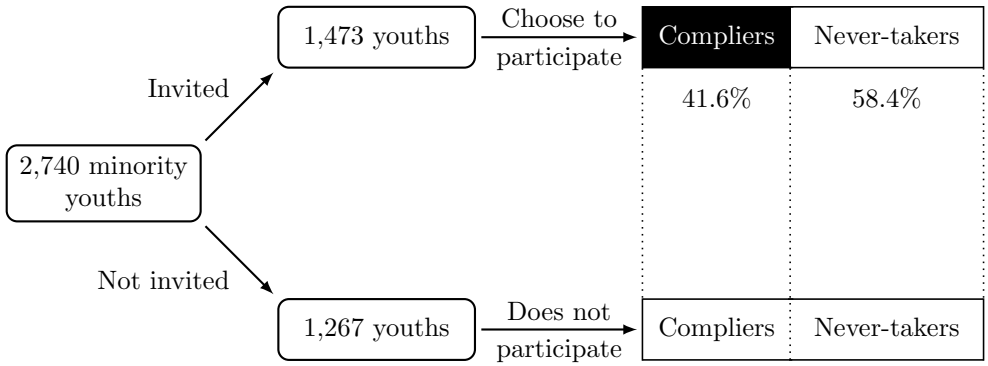


FIGURE 9.6. Randomization in the BAM experiment [HSG+17].

That is, $\alpha_1 = \pi\mathbb{E}[Y_{i1} - Y_{i0} \mid \text{complier}]$. In other words, ITT is the proportion of compliers times the average treatment effect for compliers. This treatment effect, $\beta_1 = \mathbb{E}[Y_{i1} - Y_{i0} \mid \text{complier}]$, is called the *local average treatment effect (LATE)* as it is local to the compliers. Note that π can be estimated by the proportion of compliers in the eligible population in the data, that is, $\hat{\pi} = \sum X_i / \sum Z_i$. With this, we can estimate LATE by $\hat{\beta}_1 = \hat{\alpha}_1/\hat{\pi}$. This is specifically known as the *Wald estimator*, and is a special case of the instrumental variable estimator (Section 9.5.2).

The reason why we can estimate the causal effect of the treatment, despite it being not random, is that *the treatment is effectively random for compliers*. Since compliers choose $X_i = Z_i$ and Z_i is random, if we condition on the subpopulation of compliers, it is as if the treatment was assigned randomly. However, for never-takers, there is no treatment, let alone random treatment, so it is only natural that we can estimate the treatment effect for compliers and only for compliers.

In general, there may be subjects who are not eligible for the treatment but make their way into it by some means. In such cases, we need to take into account two more groups. Those who would participate regardless of eligibility, $X_i = 1$, are called the *always-takers*, and those who would participate only when they are not eligible, $X_i = 1 - Z_i$, are called the *defiers*. This complicates the analysis very much, and is out of the scope of this course.

EXAMPLE 9.4 (Becoming a Man experiment). There is a large imbalance in criminological statistics across races. This may be due to institutional factors, or maybe due to individual choices and behaviors such as dropping out, drug uses, and entering gangs. The "Becoming a Man (BAM)" program was developed by the Chicago NGO Youth Guidance to see if we can intervene in the individual behavior aspect and improve minorities' crime participation.[3] In 2009, 2,740 minority youths were randomly assigned to the control and the treatment groups. Of them, 1,473 received an invitation to the program; 1,267 did not. Of the invited youths, 41.6% complied and participated in the program (Figure 9.6). [HSG+17] analyze this dataset to estimate the causal effect of the program on the subsequent school attendance and

[3]For an example of the program activity, see [HSG+17, p. 3].

TABLE 9.4. Becoming a Man experiment [**HSG+17**, Tables IV and VI].

Outcome	ITT		LATE	
School engagement	0.057*	(0.022)	0.137*	(0.051)
Graduated on time	0.030	(0.016)	0.071	(0.038)
Total arrests per year	−0.078	(0.046)	−0.187	(0.109)
Violent	−0.035*	(0.017)	−0.083*	(0.039)
Property	0.005	(0.013)	0.012	(0.030)
Drug	0.001	(0.018)	0.003	(0.042)
Other	−0.050	(0.027)	−0.119	(0.065)

crime statistics of these youths (Table 9.4). For example, they estimate ITT on the index for school engagement index at 0.0569, and then divides it by 0.416 to obtain LATE at 0.1367, which is found significant. They also find that participation in the program has significantly decreased the yearly arrests for violent crimes by −0.0829, which translates to a decrease by 45–50% relative to compliers who did not participate in the program.

EXAMPLE 9.5 (Vietnam draft lottery). National defense is of great importance, and there is a consensus that veterans should be adequately compensated for their service. What exactly constitutes an adequate compensation is more contentious. [**Ang90**] investigates the causal effect of attending the war on veterans' lifetime earnings. Technically, this is a causal inference with observational data, but the situation is very close to a controlled experiment. In 1970–2, men of ages 19–26 were randomly drafted based on their birthdays. Once they were drafted, they went through the physical examination and the mental aptitude test. Those who passed both were sent to Vietnam; those who failed either were not. Among those who were not eligible, some of them volunteered to enlist; of them, those who passed both tests were sent to Vietnam. This situation can be analyzed by the same method as experiments with random eligibility. To see this, observe that men of eligible ages are split into three groups: (1) those who would volunteer and could pass the tests, (2) those who would not volunteer but could pass the tests, and (3) those who could not pass the tests (Figure 9.7). The first group would go to Vietnam regardless of whether they are drafted, and the third groups would not go to Vietnam regardless. The second group, however, would go to Vietnam if and only if they are drafted. This means that their veteran status (the treatment) was assigned randomly. Therefore, a similar ratio estimator can estimate LATE for these individuals. The author first estimates ITT, the causal effect of being draft-eligible, on the income at the age of 31, to be −$487.8 for the 1950 cohort. This means that being drafted and taking the tests, regardless of passing or failing, had decreased their income 10 years later for about $500 a year. The standard error was $237.6 and it was significant. Next, the author estimates that the proportion of those who would not volunteer but could pass the tests to be 15.94%, which is the difference of the proportion of draft-eligible individuals going

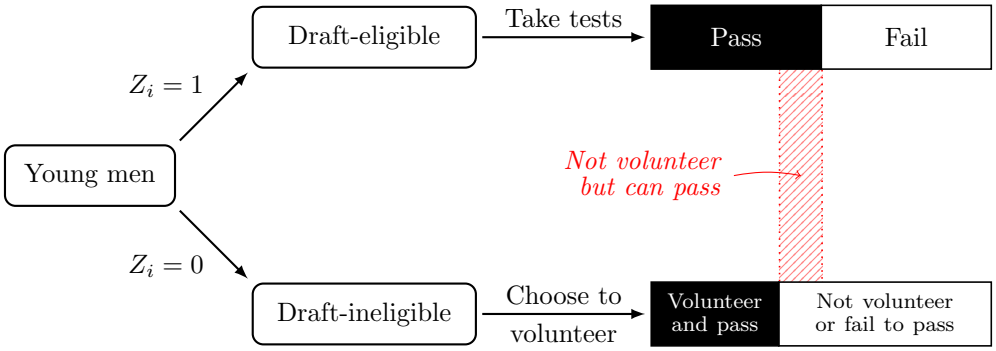


FIGURE 9.7. Drafted individuals went to Vietnam if they passed the medical exam and mental aptitude test. Draft-ineligible individuals went to Vietnam if they volunteered and passed both tests. Therefore, for those who would not have volunteered but could pass the tests, service was randomly assigned.

TABLE 9.5. Causal effect of military service on income at age 31 [**Ang90**, Tables 2 and 3].

Cohort	Year	ITT [\$/yr]	P(Pass)	P(Vol. and Pass)	LATE [\$/yr]
1950	1981	−487.8*	0.3527	0.1933	−3,060.2*
		(237.6)	(0.0325)	(0.0233)	(1,490.6)
1951	1982	−278.5	0.2831	0.1468	−2,043.3
		(264.1)	(0.0390)	(0.0180)	(1,937.6)
1952	1983	−500.0	0.2310	0.1257	−4,748.3
		(294.7)	(0.0473)	(0.0146)	(2,798.7)

to Vietnam, 35.27%, and the proportion of draft-ineligible ones going to Vietnam, 19.33%. Dividing ITT with this number, the author obtains −\$3,060.2 as the LATE of going to Vietnam on the income 10 years later, which is about 15% less than those who did not attend the war.

9.4. The Problem of Endogeneity

9.4.1. Observational data and causality. Although an experiment is the golden formula for causal inference, it is not always possible when it comes to social science; experiments may be too costly, infeasible, or unethical to run. This motivates carrying out causal inference with observational data, which can be obtained much easier and quicker with much lower costs. To develop a framework for causality in observational studies, it is essential to distinguish the two types of regression equations: predictive equations and causal equations.

Suppose that an economist is interested in uncovering the causal relationship of school budgets on graduation rates, which might provide a basis for an education

policy reform. Let y_i be the graduation rate of high school i and x_i be the budget allocated to this school. She collects the data from many districts and runs the regression

$$y_i = \gamma_0 + \gamma_1 x_i + u_i$$

and finds that $\hat{\gamma}_1$ is significantly positive. Can she conclude that increasing the budget for a school with a currently low graduation rate will boost up its graduation rate? Not really. It may be that low-budget schools have low budgets because of some underlying socioeconomic environments in the corresponding districts, which also are leading to low graduation rates. If this is the primary cause of the positive $\hat{\gamma}_1$, then allocating larger budgets on low-budget schools would be a futile effort in increasing the graduation rates. Therefore, the above equation is different from what would actually happen when we intervene in the school budget,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

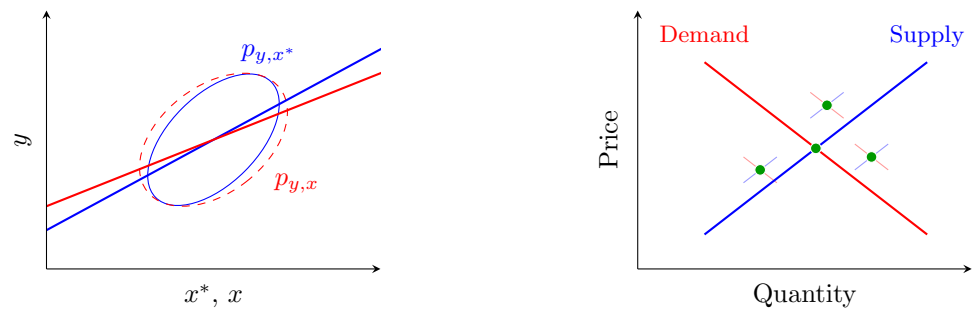
where β_1 represents the causal effect of changing the budget on the graduation rate of school i. It is important to distinguish the two, so we call the first equation the *predictive equation* and the second the *causal equation*.

In this example, the problem comes from the fact that the economist is interested in the causal equation while she estimated the predictive equation. Don't think that the predictive equation is "wrong" by any means; it just serves a different purpose. For example, if you are betting money on the graduation rate of a school randomly picked, and you are only given the information of the budget, then you would want to adjust your bet according to γ_1, not β_1. In this sense, the predictive equation is useful when *prediction from the same population* is of concern. Meanwhile, if you are negotiating the budget to improve the graduation rate, then you would want to make your argument based on β_1. In other words, the causal equation is useful when *intervention* is of concern.[4]

But then, when she ran the regression, why did she end up estimating the predictive equation and not the causal one? The key is in the data she used. Recall that the linear regression always estimates the best predictor line so that the regressors and the error term are uncorrelated in the population (Section 7.1.2). But such a predictor is only best if you use it to predict the outcome for the *same* population. Since the data she used came from an observational population, her regression line is best for predicting a new set of data that, too, comes from an observational population. Yet, her intention was in predicting the outcome of intervention, which distorts the population distribution. Mathematically put, the x_i in the observational data is uncorrelated with u_i in the predictive equation but may not be so with ε_i in the causal equation.

This problem—when we have a causal equation in mind but x_i in our dataset is correlated with ε_i in the causal equation—is called *endogeneity*. This variable x_i is called the *endogenous* variable. In turn, a variable that is uncorrelated with ε_i

[4]Intervention is not the only reason to care about causality. When a litigation consulting firm makes a case for an appropriate compensation for discrimination, the causal equation is of interest for attributive purposes.



(A) Measurement error in the regressor introduces noise in the horizontal direction, thereby making the regression line flatter (attenuation bias).

(B) Each observation is a result of the two fluctuating functions, so the line that approximates the observations is neither the demand nor the supply function.

FIGURE 9.8. Measurement error and simultaneity.

is called *exogenous*. Note that this definition depends on other regressors included in the regression; with additional regressors, an endogenous variable may become exogenous.

On the other hand, suppose that we have a dataset from an experiment in which the budgets are assigned randomly (although such an experiment would be obviously infeasible). Then, linear regression estimates the best linear predictor in that experimental population. In particular, the distribution of "everything else" is the same across all budgets. In this case, the predictive equation for this population coincides with the causal equation, and she would end up estimating the causal effect.

The goal of causal inference with observational data is, therefore, to find the variation in the observational data that is *as if* an experiment is carried out, whereby the predictive equation coincides with the causal equation.

9.4.2. Sources of endogeneity. Endogeneity can arise in various ways. Below are some notable channels. Note that all these are mathematically equivalent, and some authors use "omitted variable bias" as a synonym for endogeneity.

EXAMPLE 9.6 (Omitted variable bias). The *omitted variable bias* takes place when there is a variable w that causes both x and y, and you miss w in the regression (recall Figure 7.9). In marketing, there is a myth that growing the market share is key to profitability [**NM18**, Chapter 7]. This is based on the observation that there is a positive correlation between the market share x and profitability y, that is, β_1 in $y = \beta_0 + \beta_1 x + \varepsilon$ is positive. However, some research shows that this relation is not causal, and that some underlying factors such as a competitive advantage w lead to both the high market share and profitability; when we run $y = \beta_0 + \beta_1 x + \beta_2 w + \varepsilon$, the coefficient β_1 is hardly positive. Thus, aiming solely for the market share (and not improving the competitive advantage) hardly contributes to profitability.

Example 9.7 (Measurement error). It has long been known that the responses in economic surveys contain some extent of errors [**MS83**]. For example, work experience may be measured in years but not in months or days; self-reported income may or may not include taxes. This is known as the *measurement error* or the *errors in variables*. In particular, suppose that the equation of interest is $y = \beta_0 + \beta_1 x^* + \varepsilon$, but x^* is not measured; instead, we measure $x = x^* + u$ with some noise u. Then, γ_1 in the regression $y = \gamma_0 + \gamma_1 x + \varepsilon$ is known to underestimate β_1 toward the origin (Figure 9.8A). This is known as *regression dilution* or the *attenuation bias* [**Gre18**, Section 8.8]. This is similar to the omitted variable bias in the sense that x^* is not observed, but in this case the causal effect of interest is $x^* \to y$. Since the bias is toward zero, if $\hat{\gamma}_1$ is significant, then it is a stronger evidence against $H_0 : \beta_1 = 0$.

Example 9.8 (Simultaneity). *Simultaneity* is a situation in which x causes y and y causes x. Demand estimation is a good example of this [**Hay00**, Section 3.1]. The demand function is a map from the price to the quantity demanded. However, the *observed* price and quantity are the intersection of the demand curve and the supply curve, so regressing the observed quantity on the observed price estimates neither the demand nor the supply function (Figure 9.8B). In fact, the instrumental variable method to be explained in Section 9.5.2 was originally invented to estimate the demand and supply curves properly [**ST03**].

Example 9.9 (Reverse causality). The *reverse causality* is when y causes x. It is sometimes believed that formula-fed infants grow more rapidly than breastfed infants. [**KMDP11**] document that infant feeding choices are affected by the infants' health; small infants are more likely to be subsequently weaned or to have discontinued exclusive breastfeeding, and large infants are less likely to experience these feeding changes. If small infants are more likely to experience natural rapid growth, the causality of feeding choices and the growth speed may be reverse.

Example 9.10 (Self-selection). *Self-selection* occurs when a particular combination of y and x is selectively observed. For example, individual workers decide on their occupations taking into account their own skill sets and fits. So, if a pianist switches to an accountant, it is not reasonable to foresee a salary raise equal to the difference of the average income of an accountant and that of a pianist; that is, the observed relationship of average income and occupation is not causal. This problem is explicitly discussed by [**Roy51**], and the economic model that describes this selection process is known as the *Roy model* in labor economics.

9.4.3. Is there "included variable bias?" As discussed in Section 9.3.1, including additional covariates uncorrelated with the current regressors x_i does not change the coefficient. Meanwhile, including a variable that is correlated with x_i would change the coefficient on x_i. In economics, causality is considered on a *ceteris paribus* basis—what happens (or would have happened) had x_i been altered but all other factors unchanged—so if there is any covariate that is correlated with x_i, we would rather want to include it. In this sense, the "included variable bias" is not an issue.

(A) Control variable approach shaves off the problematic part w from ε by including it in the regression.

(B) Instrumental variable approach shaves off the problematic part from x by replacing x with $\hat{x}$ fitted by z.

FIGURE 9.9. Two solutions to endogeneity. The problematic part is the intersection of x and causal ε.

In law, on the other hand, we may occasionally be interested in discovering an indirect (confounding) effect on a *mutatis mutandis* basis—how the change in x_i affects other variables that further affect y_i. For example, the *disparate impact* studies examine how race-neutral policies affect different races disproportionately through various channels [**Ayr05**]. In such cases, the included variable bias can be a major concern.

9.5. Causal Inference with Observational Data

There are two solutions to the endogeneity problem. Write the causal equation of interest as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where β_1 captures the causal effect of x_i on y_i. The problem is that in the observational data, the variable x_i is correlated with ε_i, that is, $\mathbb{E}[x_i\varepsilon_i] \neq 0$. This ε_i can be decomposed into two parts: the component that is correlated with x_i and the remainder. The correlated component is the problematic part, and is called the *confounder*. Some confounders may be observable while some others are not. The omitted variables are examples of unobservable confounders.

The first solution, which we call the control variable approach, is to shave off the error term ε_i so that it becomes uncorrelated with x_i (Figure 9.9A). This is done by adding the confounding variables that absorb the endogeneity altogether. This solution is possible when there is no unobservable confounder. Likewise, x_i can be decomposed into the part that is correlated with ε_i and the remainder. The second solution, the instrumental variable approach, is to scrape off x_i so that it becomes uncorrelated with ε_i (Figure 9.9B). This is done by projecting x_i onto an external variation z_i that is only correlated with x_i. This solution is viable when we observe a "pure" variation of x_i that is uncorrelated with ε_i. In the end, we will see that these approaches are mere flip sides of the same coin.

9.5.1. Control variable approach. Suppose we are interested in the causal effect of attending a charter school on academic achievements. From the education policy perspective, an unbiased assessment of charter schools is important as they are publicly funded and granted a large extent of freedom in education. Let y_i be the test

score of student i, used as a proxy for academic achievements, and x_i be the number of years student i has attended a charter school. If we retrieve the observational data and run the regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

then β_1 captures not only the causal effect of charter schools but all other factors that produce correlation between charter attendance and the test score. For example, the parents who think of sending their children to charter schools may be inherently education-conscious, and their children may have gained more educational opportunities outside of school. Then, it creates positive correlation between y_i and x_i even in the absence of the causal effect of attending a charter school. Also, some public schools serve a student body of disproportionately more minorities, and some of these minorities may be in disadvantaged environments. Then, it may also add to non-causal correlation between y_i and x_i.

If we observe these variables, including them in the regression changes the coefficient of x_i. Suppose that we observe how education-conscious the parents of i are and whether i is a minority; let w_i denote a vector of these variables. If we run

$$y_i = \beta_0 + \beta_1 x_i + w_i' \beta_2 + \varepsilon_i,$$

then β_1 is the (causal and non-causal) effect of x_i on y_i *fixing the variable* w_i *at the same level* (recall the discussion in Section 7.1.2). In other words, β_1 is the average of the differences of students in charter schools and in other public schools in which the comparison is made minority to minority, majority to majority, education-conscious parents to education-conscious parents, etc. Thus, if we include all of those non-causal channels in the regression (which refines the error term), we will eventually be left with the pure causal effect of x_i in β_1. This is the idea behind the control variable approach. The vector w_i is the set of *control variables*, and this regression is said to *control for* w_i.

As is clear from the discussion, the challenge of this approach is that we need to observe *all* sources of endogeneity and need to know the correct functional form of how these covariates affect y_i. The assumption that we observe the entire confounding factor is called the *selection on observables* or *conditional independence* assumption. On the flip side, an advantage of this method is that what is held fixed is very clear; we control for the covariates and only the covariates.

[**AAD+11**, Table X] use this approach to estimate the effect of a charter school on academic achievements. Let $\mathit{TestScore}_i$ be the test score of student i on either English or Math and $\mathit{YearsInCharter}_i$ be the time student i has spent in a charter school measured in years. Then, they run the regression [**AAD+11**, eq. (11)]

$$\mathit{TestScore}_i = \beta_0 + \beta_1 \mathit{YearsInCharter}_i + w_i' \beta_2 + \varepsilon_i,$$

where w_i is a vector of covariates that includes demographic controls such as ethnicity, gender, English proficiency, whether the student is receiving free/reduced price lunch, the interaction of gender and ethnicity, and the baseline test scores before attending (or choosing not to attend) the charter schools.

TABLE 9.6. The effect of charter schools estimated by the control variable approach [AAD+11, Table X]. The test scores are measured in the units of standard deviations.

	Outcome	
	English [σ]	Math [σ]
Years in charter	0.174*	0.316*
	(0.020)	(0.024)
N	40,852	45,035

In reference to the above discussion, the baseline test scores may control for the education-consciousness of the parents since such parents would have given their children additional education opportunities well before sending them to charter schools.

Table 9.6 gives the estimates of β_1 for English and Math. If we buy that these control variables are the only possible channels that can induce non-causal correlation between the test scores and charter attendance, we can interpret Table 9.6 as the evidence that one year of attendance to a charter school increased the English test scores by 0.174 standard deviations and the math scores by 0.316 standard deviations on average for the students included in the regression.

On the other hand, if we believe that there are also other channels that can induce non-causal correlation, we can interpret that the coefficients are the total effects including these channels. For example, if we believe that early childhood education affects not only the level but also the speed at which children learn things in later ages, then having students with more preschool education at charter schools than at other public schools can create a spurious positive correlation between charter attendance and test scores. In this case, 0.174 refers to the causal effect of attending a charter school plus the effect of having more preschool education on the English test scores.

EXAMPLE 9.11 (Search engine marketing). Search engines sell their ad spaces based on search keywords. It is of interest to potential advertisers to know which keywords they should buy to maximize the ad effect. Particularly, eBay used to pay for the ads when users searched for keywords containing "eBay." As a result, a user who searched for keywords such as "eBay shoes" saw an ad that led to eBay's website as well as the natural search result that also led to their website. Is it meaningful to show the ad when the users' intention seems already to reach their website? To see this, eBay stopped paying MSN (Bing) for the ad for "eBay"-containing keywords in March 2012, while they kept the ad on Google. [BNT15] estimate the causal effect of terminating the ads for brand names using the *difference-in-differences (DID)* method. To understand this method, divide the data into four quadrants: (1) Google vs. Bing and (2) before vs. after halting the ads (Figure 9.10). First, consider the naive regression $y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$, where y_{it} is the logarithm of the number of clicks originating from the search engine i at time t, separately for each search engine

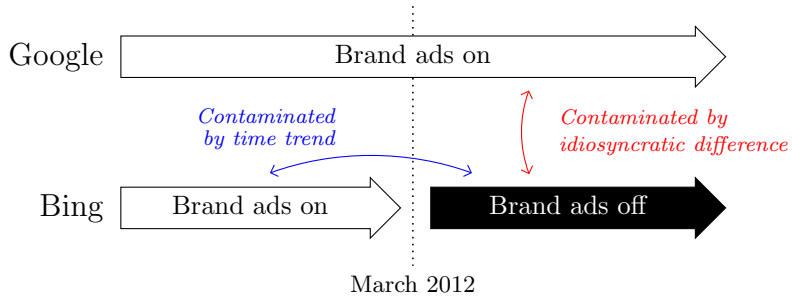


FIGURE 9.10. Comparison between Google and Bing after March 2012 captures unwanted idiosyncratic difference. Comparison of Bing before and after March 2012 captures unwanted time trend. Difference-in-differences eliminates this contamination under parallel trend.

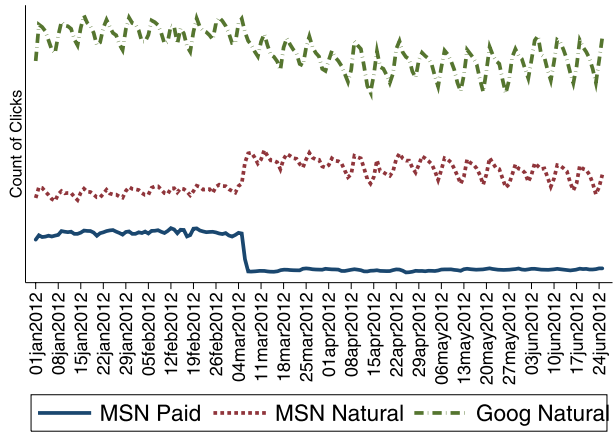


FIGURE 9.11. Click-traffic counts for Google and MSN Bing for the duration of the experiment [**BNT15**, Figure 2]. "Natural" refers to clicks through natural search results, and "Paid" through ads.

(Table 9.7). The authors find a significant negative effect for Bing, $\hat{\beta}_1 = -0.056$, which means that eBay experienced a significant drop in its traffic after halting the ads. However, the same regression for Google also yields a significant negative effect, $\hat{\beta}_1 = -0.032$. This indicates that the negative effect for Bing might not be due to stopping the ads, but something else that affected the overall traffic for eBay. The idea behind DID is to adjust Bing's traffic using Google's. If some eBay-specific event happened that decreased the overall traffic for eBay, then what happened to Google's traffic (traffic difference $y_{12} - y_{11}$) may be what would have happened to Bing's traffic had eBay kept the ads on Bing (*parallel trend assumption*). Then, the traffic that Bing would have experienced had it kept the ads can be estimated as $y_{21} + (y_{12} - y_{11})$. Then, the actual traffic minus this counterfactual difference, $(y_{22} - y_{21}) - (y_{12} - y_{11})$, is

TABLE 9.7. Naive regression and difference-in-differences [BNT15, Table A.I]. The outcome variable is the log click counts. The indicator After is replaced with various fixed effects for DID.

	Google	Bing	
	Naive	Naive	DID
After	−0.032*	−0.056*	—
	(0.012)	(0.009)	—
Ads off			−0.005
			(0.018)
Google			5.088
			(10.06)
Yahoo!			1.38
			(5.66)
Constant	14.34	12.82	11.33
	(0.006)	(0.006)	(5.66)
N	120	118	180

the estimated traffic on Bing that is due to halting the ads.[5] This estimator takes the form of taking a difference of the two differences, hence the name. Mathematically, this is equivalent to running the regression

$$y_{it} = \beta_0 + \beta_1 AdsOff_i + \beta_2 After_i + \beta_3 Google_i + \varepsilon_{it}.$$

The terms $\beta_2 After_i$ and $\beta_3 Google_i$ are called the *fixed effects*, and absorbs the idiosyncratic effects for the post-experiment period and for Google, respectively. The parallel trend assumption implies that these fixed effects capture all sources of endogeneity regarding this ad effect. The actual movements of log clicks are shown in Figure 9.11. The authors find that the ad effect is now insignificant at $\hat{\beta}_1 = -0.005$,[6] concluding that "the evidence strongly supports the intuitive notion that for brand keywords, natural search is close to a perfect substitute for paid search, making brand keyword SEM ineffective for short-term sales." eBay subsequently stopped paying for brand keywords on other platforms as well.

EXAMPLE 9.12 (The Fox News effect). According to the measure constructed by [GM05], most media outlets are biased in some sense (Figure 9.12). Then, does media bias affect voting? [DK07] use the introduction of Fox News to estimate the effect of the availability of Fox News on conservative votes. Since Rupert Murdoch launched Fox News in 1996, it has acquired as much audience as CNN by 2000. The

[5]Unless we assume that what had happened to Bing at $t = 2$ is what would have happened to Google had they halted the ads on Google, we cannot say that this estimate applies also to Google. In this sense, the target of the DID estimator is called the *average treatment effect on the treated (ATT)*. Here, Bing is the "treated" and Google the "control."

[6]The actual regression carried out in Table 9.7 includes more fixed effects and data on Yahoo!

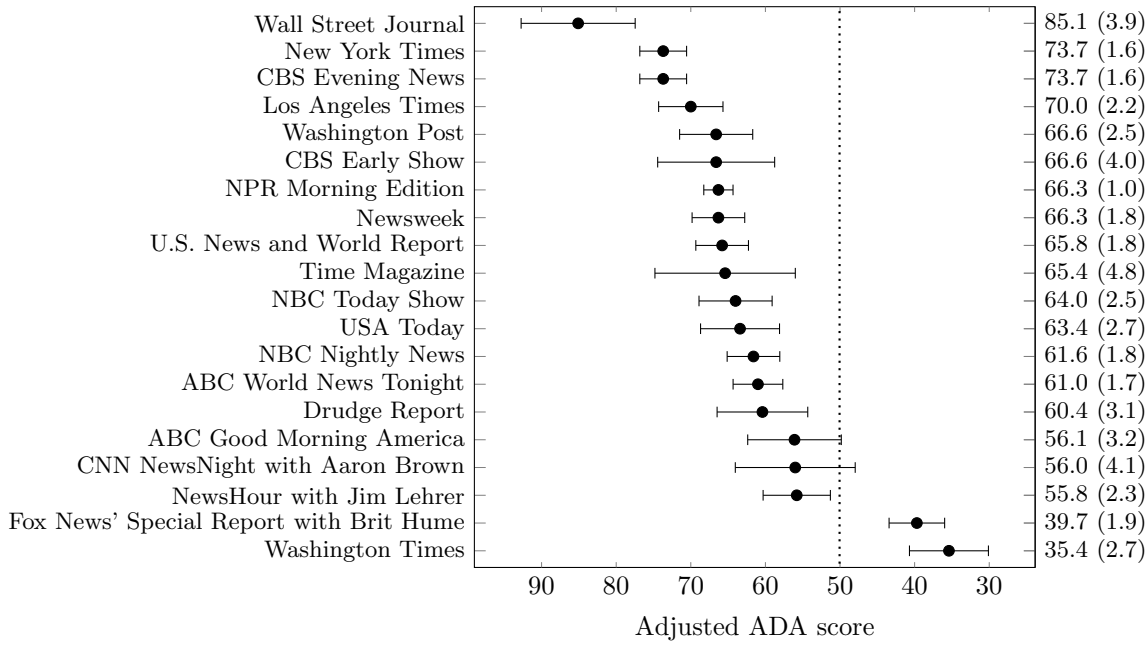


FIGURE 9.12. The measure of media bias "adjusted Americans for Democratic Action (ADA) score" calculated by [**GM05**]. The error bars indicate marginal 95% confidence intervals assuming normality. The dotted line is the political center defined by the authors.

authors argue that "Fox News availability in 2000 appears to be largely idiosyncratic, conditional on a set of controls." From there, they apply the difference-in-differences methodology to estimate how Fox News shifted the conservative vote share in Presidential elections between 1996 and 2000. Their regression equation is

$$GOPVoteShare_i = \beta_0 + \beta_1 Fox_i + \beta_2 2000_i + W_i'\beta_3 + \varepsilon_i,$$

where $GOPVoteShare_i$ is the Republican vote share in town i, Fox_i is the indicator of whether Fox News was available in town i, 2000_i is the indicator of the election in 2000, and W_i is a set of controls including the demographic variables and the features of the cable system. They estimated $\hat{\beta}_1 = 0.0069$ with the standard error 0.0014, which is significant. Thus, the Republican candidate is considered to have gained 0.7% of votes due to Fox News in the available towns.

EXAMPLE 9.13 (Value of connections in turbulent times). On November 21, 2008, the news leaked that Timothy Geithner would be nominated as the U.S. Treasury Secretary for President Obama. In the next several days, the stock prices of firms that have connections with Geithner, such as JPMorgan Chase, have surged relative to the stocks of firms with no connections, such as Charles Schwab. On January 13, 2009, Geithner's unexpected tax issues were exposed. Subsequently, the Geithner-connected firms' stocks plunged (Figure 9.13). [**AJK+16**] use this to measure the "value" of having a connection to the U.S. Treasury Secretary in the time of a crisis.

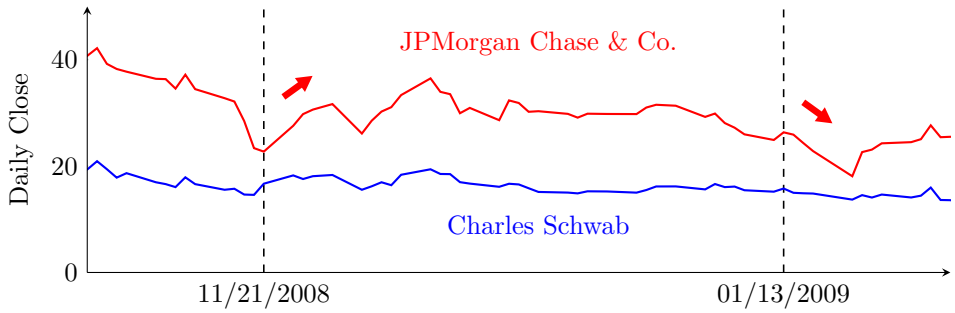


FIGURE 9.13. The stock prices of JPMorgan Chase & Co. and Charles Schwab. On November 21, 2008, the news leaked that Timothy Geithner would be nominated as the U.S. Treasury Secretary. On January 13, 2009, Geithner's unexpected tax issues were exposed.

They consider the regression

$$AbnormalReturns_i = \beta_0 + \beta_1 GeithnerConnections_i + W_i'\beta_2 + \varepsilon_i$$

in the span of ten days before and after the events. The control variable W_i includes firm characteristics such as the firm size, profitability, and leverage. Assuming that the news were unexpected to the market and W_i absorbs imbalance between firms with and without connections, β_1 measures the causal effect of having connections to the U.S. Treasury Secretary on the stock returns of a financial institution during a financial crisis. The authors find that $\hat{\beta}_1$ is significant and positive at 0.0073.

9.5.2. Instrumental variable approach. The instrumental variable method is a generalization of the Wald estimator we saw in Section 9.3.2. Recall that the Wald estimator finds the subpopulation for which the assignment of the treatment was random. Similarly for the observational data, we look for the variation that randomly changed the x_i of which we want the causal effect on y_i.

Let us continue on the working example of the charter school effect. Many charter schools receive more applications than they can admit, and as a result, they hold a lottery. Then, the lottery is a variation that affects the charter attendance x_i but does not affect other confounding variables; e.g., the minority status of the students or the wealthiness of their parents. Therefore, among those students who went through the lottery—be they admitted or not—the situation is *as if* an experiment was conducted and they were randomly admitted. This type of solution is possible when there is an external variation z, called the *instrument* or the *instrumental variable*, that is only correlated with x (Figure 9.14).

In the instrumental variable method, we first extract the variation of x_i that came from such an exogenous shock only. This is done by the following regression, called the *first-stage equation*,

$$x_i = \pi_0 + \pi_1 z_i + v_i,$$

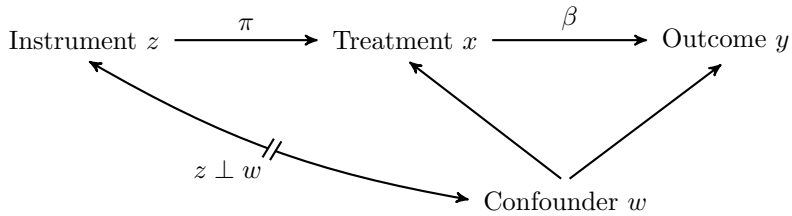


FIGURE 9.14. Causal graph for the instrumental variable method. Even if w is unobservable, the causal effect β of x on y can be found if there is a variable z that affects x but not w.

where z_i is the indicator of whether the student i won the lottery. With this, we can "predict" x_i using only the information of z_i by

$$\hat{x}_i \coloneqq \hat{\pi}_0 + \hat{\pi}_1 z_i.$$

While this is clearly correlated with x_i, this is not correlated with the confounding variables since this prediction only uses the information of z_i, which has no implications on variables other than x_i. Mathematically, substitute $x_i = \hat{x}_i + \hat{v}_i$ into the original (*second-stage*) equation

$$y_i = \beta_0 + \beta_1(\hat{x}_i + \hat{v}_i) + \varepsilon_i = \beta_0 + \beta_1\hat{x}_i + (\beta_1\hat{v}_i + \varepsilon_i).$$

The last term in the parentheses is the padded error term as a result of substituting x_i, and is uncorrelated with z_i and hence with $\hat{x}_i$. Therefore, by regressing y_i on $\hat{x}_i$, we find an estimator $\hat{\beta}_1$ of our target causal effect. The condition for an instrument to be valid is, therefore, that it is correlated with x_i (so that $\hat{x}_i$ is meaningful at all) and it is not correlated with ε_i. The first condition is verifiable while the second is not.[7]

In sum, the instrumental variable method proceeds in two steps. First, regress the endogenous regressor x_i on the instrument z_i. Second, regress the dependent variable y_i on the fitted regressor $\hat{x}_i$. The estimator $\hat{\beta}_1$ thusly constructed is called the *two-stage least squares (2SLS or TSLS)* estimator. Note that we can also include other covariates in this method; just make sure to include them in *both* of the first- and second-stage equations.

[**AAD+11**, Table IV] estimate the effect of a charter school using this method (Table 9.8). As stated above, they use the lottery as an instrument for the time spent in charter schools. They find the causal effect estimates $\hat{\beta}_1$ quite close to the ones found by the control variable approach and state that "the observational study design does a good job of controlling for selection bias in the evaluation of charter effects (or that there is not much selection bias in the first place)." It is, however, rare that the same effect be estimated using both methods. In many cases, we have some okayish instruments and not-too-many control variables, so we combine the two to boost credibility.

[7] If there is more than one instrument for one endogenous variable, we can test if any one of the instruments is invalid.

TABLE 9.8. The effect of charter schools estimated by the two approaches [AAD+11, Tables IV and X]. The test scores are measured in the units of standard deviations.

	Control variable		Instrumental variable	
	English [σ]	Math [σ]	English [σ]	Math [σ]
Years in charter	0.174*	0.316*	0.198*	0.359*
	(0.020)	(0.024)	(0.047)	(0.048)
N	40,852	45,035	3,101	3,258

The advantage of the instrumental variable approach is that we need only one instrument for one endogenous regressor. This is in contrast to the control variable approach requiring *all* sources of endogeneity be included. However, the major disadvantage, too, is that we have to find at least *one* instrument. It is indeed quite challenging to find even one valid instrument that is convincing to many. For this reason, finding a neat instrument can sometimes even change the course of the literature; in economics, there are a few "well-known" instruments that are given names and frequently referred to, e.g., the BLP instrument, the Bartik instrument, the twin birth instrument, and the quarter-of-birth instrument.

EXAMPLE 9.14 (Rainfall for demand estimation). One of the most famous examples of an instrument in economics is rainfall. As discussed in Example 9.8, estimation of the demand curve suffers from the endogeneity of the price. Consider the demand estimation of coffee beans in the United States. The quantity and price of the coffee beans is determined by a system of demand and supply equations in the U.S., so a simple regression of the quantity on the price does not find the demand curve or the supply curve. However, most coffee beans are produced outside of the country, e.g., in South America. Thus, the rainfall in the coffee production locations in South America would affect the supply curve while staying independent of the demand for coffee in the U.S. From a standpoint of the demand curve, therefore, the rainfall in South America acts as an instrument that affects the price x_i but not the confounding factors ε_i; thereby, it is *as if* the price-changing experiment was carried out randomly to North American consumers. It may at first seem somewhat counterintuitive that, in order to estimate the demand curve, we need a variation that does *not* affect the demand curve. However, pondering over the intuition above, it should eventually come natural to you.

EXAMPLE 9.15 (Class size and test scores). There is an argument that education is more effective in small class sizes as teachers can pay more attention to each student. However, its implementation is costly to the public and hence requires scientific evidence. Therefore, investigating whether and how much smaller class sizes help improve education quality is an important economic question to answer. Medieval philosopher Maimonides once said, "If there are more than 40 [children], two teachers must be appointed." Israel keeps up to his words and sets the maximum of 40 pupils

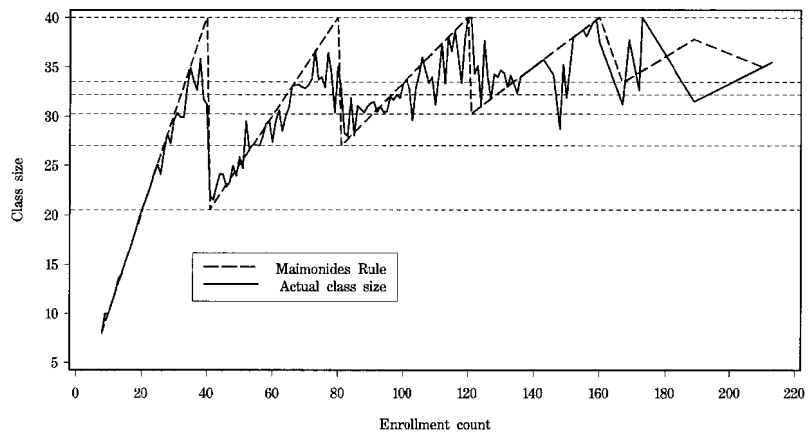


FIGURE 9.15. Maimonides's rule and actual class size [**AL99**, Figure I].

per class in public schools (Figure 9.15). Continuing on Example 7.5, [**AL99**] use this to estimate the causal effect of class size on test scores. They argue that the class size predicted by Maimonides's rule is a valid instrument since most children in Israel attend public schools and it is hard to predict which district ends up in small class sizes. The first-stage equation is

$$ClassSize_i = \pi_0 + \pi_1 MaimonidesRule_i + W_i'\pi_2 + v_i,$$

where $ClassSize_i$ is the actual size of class i, $MaimonidesRule_i$ is the size predicted by Maimonides's rule, and W_i is the vector of shool-level covariates including the percentage of pupils with disadvantaged backgrounds. The second-stage equation is

$$Reading_i = \beta_0 + \beta_1 ClassSize_i + W_i'\beta_2 + \varepsilon_i,$$

where $Reading_i$ is the average test score for class i. [**AL99**, Table IV] calculate that the 2SLS estimate $\hat{\beta}_1$ is -0.410 with standard error 0.113, now much larger than -0.053 in Example 7.5. They conclude that "[t]he raw positive correlation between achievement and class size is clearly an artifact of the association between smaller classes and the proportion of pupils from disadvantaged backgrounds." "The [instrumental variable] estimates show that reducing class size induces a significant and substantial increase in test scores for fourth and fifth graders, although not for third graders."

EXERCISE 9.1 (Causal effect of studying). If I study for one more hour, how much does my GPA improve? This is an existential concern for many students. Let y_i be the GPA of student i and x_i be the average number of hours i studies per day. If we run the regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, can we interpret β_1 as the causal effect of studying? If not, what confounding effects does β_1 capture? [**SS08**] take advantage of a unique dataset to overcome this endogeneity problem. Let z_i be the indicator of whether i's roommate brought a video game. The authors propose to use z_i as an instrument for x_i. What conditions does z_i need to satisfy in order for this to work? At Berea College, students are assigned to dorms randomly; this means that students have no control over who their roommates will be except for their genders.

Is this important in the validity of z_i? What other control variables do you suggest including?

9.5.3. Equivalence of the two approaches. The control variable approach and the instrumental variable approach are two different methods that serve conceptually distinct situations. The first can be used when we observe granular data that cover all sources of endogeneity; the second when we observe a delicate and peculiar variable that affects the endogenous variable but none of the confounding variables.

Mathematically, however, the two approaches are equivalent; the 2SLS estimator can be obtained by regressing y on x and the residual from the first stage, that is,

$$y_i = x_i'\beta + \hat{v}_i'\gamma + \varepsilon_i.$$

To see this, let $X = Z\hat{\pi} + \hat{V}$ be the estimated first-stage equation and consider $\hat{\beta}$ from the following regression

$$Y = X\beta + \hat{V}\gamma + \mathcal{E}.$$

By Theorem 7.11, $\hat{\beta}$ satisfies

$$\hat{\beta} = (X'M_{\hat{V}}X)^{-1}X'M_{\hat{V}}Y = [(M_{\hat{V}}X)'(M_{\hat{V}}X)]^{-1}(M_{\hat{V}}X)'Y$$

for idempotent $M_{\hat{V}} = I - \hat{V}(\hat{V}'\hat{V})^{-1}\hat{V}'$. Observe that

$$M_{\hat{V}}X = [I - \hat{V}(\hat{V}'\hat{V})^{-1}\hat{V}'](Z\hat{\pi} + \hat{V}) = Z\hat{\pi} = \hat{X}$$

since $\hat{V}'Z = 0$. Therefore, we obtain

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y,$$

which is numerically identical to the 2SLS estimator. Nevertheless, the estimator thusly obtained is sometimes distinguished by the name *two-stage residual inclusion (2SRI)* estimator to highlight the conceptual difference. One advantage of the 2SRI formulation is that it is easily generalizable to nonlinear models such as quantile regression. Such generalization is called the *control function* method.

This exercise reveals that the variation of x_i *orthogonal to* z_i acts as the control variable that absorbs all the confounding effects. So, at the end of the day, there is only one solution to get to the causal effects inductively—we have to identify all sources of endogeneity, either by directly observing them or by observing a variation orthogonal to them.

9.A. Theory of Two-Stage Least Squares

Let x_i be the k-dimensional vector of regressors, which includes both endogenous and exogenous regressors. An intercept, if included, is by construction exogenous. Let z_i be the ℓ-dimensional vector of instruments, which includes the instruments for the endogenous part of x_i as well as all exogenous regressors of x_i.[8] There must be at least as many instruments as endogenous regressors, so $k \leq \ell$.

[8]So, the exogenous regressors act as their own instruments.

The linear IV model is defined by the two equations

$$\begin{cases} y_i = x_i'\beta + \varepsilon_i, \\ x_i' = z_i'\pi + v_i', \end{cases}$$

where the first equation is called the *second-stage equation* and the second the *first-stage equation*. The substituted equation $y_i = z_i'\pi\beta + u_i$, for $u_i = \varepsilon_i + v_i'\beta$, is called the *reduced-form equation*. Note that v_i is a $k \times 1$ vector as in x_i and π is an $\ell \times k$ matrix, so the first-stage equation is a collection of k equations. If z_i contains a subvector of x_i, then the corresponding part of the first-stage equation is redundant but innocuous. Denote the dataset we have as

$$\underset{(n\times 1)}{Y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{(n\times k)}{X} := \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}, \quad \underset{(n\times \ell)}{Z} := \begin{bmatrix} z_1' \\ \vdots \\ z_n' \end{bmatrix} = \begin{bmatrix} z_{11} & \cdots & z_{1\ell} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{n\ell} \end{bmatrix}.$$

The 2SLS estimation identifies the parameter (π, β) that satisfies $\mathbb{E}[z_i v_i] = 0$ and $\mathbb{E}[z_i \varepsilon_i] = 0$ and proceeds in two steps.

(1) Regress x on z to obtain $\hat{\pi} = (Z'Z)^{-1}Z'X$. Calculate the fitted values of x, that is, $\hat{X} = Z\hat{\pi} = Z(Z'Z)^{-1}Z'X$.
(2) Regress y on $\hat{x}$ to obtain $\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$.

The OLS estimator is a special case of this in which all regressors are assumed exogenous, i.e., $z_i = x_i$.

9.A.1. Statistical properties. The 2SLS estimation has two essential assumptions (Assumptions 9.1 and 9.2).

Assumption 9.1 (Sampling assumption)**.** $\{y_i, x_i, z_i\}$ are i.i.d. with second moments and $k \le \ell$. $\mathbb{E}[z_i z_i']$ is invertible, $\mathbb{E}[z_i x_i']$ is of full row rank, and $\mathbb{E}[\varepsilon_i^2 z_i z_i']$ exists.

Assumption 9.2 (Instrumental orthogonality)**.** There exists β such that $\mathbb{E}[(y_i - x_i'\beta)z_i] = 0$.

Remark 9.1. Assumption 9.2 is trivial if x and z have the same dimension.

The next assumption is not necessary but can simplify the variance formula.

Assumption 9.3 (Unconditional homoskedasticity)**.** $\mathbb{E}[\varepsilon_i^2 z_i z_i'] = \mathbb{E}[\varepsilon_i^2]\mathbb{E}[z_i z_i']$.

There are two ways to impose conditional versions of the assumptions to yield finite-sample distributions. One is to condition on $\hat{X}$ and the other on Z. We will include the corresponding assumptions in the statements as necessary.

The first-stage coefficient is $\pi := \mathbb{E}[z_i z_i']^{-1}\mathbb{E}[z_i x_i']$. Mathematically, the parameter of interest is defined as $\beta := (\pi'\mathbb{E}[z_i z_i']\pi)^{-1}(\pi'\mathbb{E}[z_i y_i])$. The estimator for π is simply the OLS estimator $\hat{\pi} := (Z'Z)^{-1}(Z'X)$, and the fitted X is given by $\hat{X} = Z\hat{\pi}$. Then, the 2SLS estimator for β is the OLS of Y on $\hat{X}$, i.e., $\hat{\beta} := (\hat{X}'\hat{X})^{-1}(\hat{X}'Y) = [X'Z(Z'Z)^{-1}Z'X]^{-1}[X'Z(Z'Z)^{-1}Z'Y]$. When $k = \ell$, it reduces to $\hat{\beta} = (Z'X)^{-1}(Z'Y)$, which is sometimes called the *IV estimator*.

Theorem 9.1 (Asymptotic normality)**.** *Under Assumptions 9.1 and 9.2, $\hat{\beta}$ is consistent to β and*

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow N\Big(0,\ (\pi'\mathbb{E}[z_i z_i']\pi)^{-1}(\pi'\mathbb{E}[\varepsilon_i^2 z_i z_i']\pi)(\pi'\mathbb{E}[z_i z_i']\pi)^{-1}\Big).$$

Under Assumption 9.3, the asymptotic variance reduces to $\mathbb{E}[\varepsilon_i^2](\pi'\mathbb{E}[z_i z_i']\pi)^{-1}$.

Proof. Since $\mathbb{E}[zz']$ is invertible and $\mathbb{E}[zx']$ is of full row rank, $Z'Z$ is invertible and $Z'X$ is of full row rank with probability approaching 1. Observe that $\hat{\beta} - \beta = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'\mathcal{E}$. Then the result follows by the LLN, CMT, CLT (applied to $\frac{1}{\sqrt{n}}Z'\mathcal{E}$), and Slutsky's lemma. The last claim is trivial. ■

Theorem 9.2 (Semiparametric efficiency)**.** *Under Assumptions 9.1 and 9.2 and either $k = \ell$ or Assumption 9.3, $\hat{\beta}$ is semiparametrically efficient.*

Proof. Apply a similar argument as [**vdV98**, Example 25.28]. ■

Remark 9.2. Just as Remark 7.3, the validity of Theorem 9.2 depends on what assumptions *not* to impose.

Under strong assumptions, we can derive the finite-sample distribution of 2SLS. There are a few ways to do so and there seems no consensus as to which is canonical. The first way is to fix the result of the first-stage regression and consider 2SLS as a mere OLS of y on $\hat{x}$. This is arguably the easiest way to get to the finite-sample distribution, but it requires us to condition on the in-sample fitted regressors, whose meaning is not clear in observational studies.

Proposition 9.3 (Finite-sample distribution conditional on $\hat{X}$)**.** *Suppose that $\mathcal{E} \mid \hat{X} \sim N(0, \sigma^2 I)$. Then, $\hat{\beta} \mid \hat{X} \sim N(\beta, \sigma^2(\hat{X}'\hat{X})^{-1})$.*

Proof. The claim is trivial given $\hat{\beta} = \beta + (\hat{X}'\hat{X})^{-1}\hat{X}'\mathcal{E}$. ■

Remark 9.3. Unlike Proposition 7.3, we do not have $\mathrm{Cov}(\hat{\beta}, \hat{\mathcal{E}} \mid \hat{X}) = 0$ unless $\hat{V} = 0$. Therefore, this finite-sample distribution does not lead to exact finite-sample testing. However, major statistical software computes critical values as if the counterpart of Proposition 7.9 holds.

We can also derive the finite-sample distribution conditional on Z by imposing normality on the reduced-form errors (see also [**Phi09**]). This way we can take into account the uncertainty from the first-stage estimation.

Proposition 9.4 (Finite-sample distribution conditional on Z)**.** *For $y_i = z_i'\pi\beta + u_i$, suppose that $(u_i, v_i')' \mid Z$ follows i.i.d. $N(0, \Sigma)$. Then, $\hat{\beta} = (\tilde{\pi}'\tilde{\pi})^{-1}(\tilde{\pi}'\tilde{\gamma})$ for which*

$$\begin{bmatrix} \tilde{\gamma} \\ \mathrm{vec}(\tilde{\pi}) \end{bmatrix} \Bigg| \, Z \sim N\left(\begin{bmatrix} \gamma \\ \mathrm{vec}(\pi) \end{bmatrix}, \Sigma \otimes I_\ell \right),$$

where I_ℓ is an $\ell \times \ell$ identity matrix.

PROOF. Note that $\hat{\beta} = [\hat{\pi}'(Z'Z)\hat{\pi}]^{-1}[\hat{\pi}'(Z'Z)\hat{\gamma}]$ where $\hat{\pi} = (Z'Z)^{-1}(Z'X)$ and $\hat{\gamma} = (Z'Z)^{-1}(Z'Y)$ follow

$$\begin{bmatrix} \hat{\gamma} \\ \mathrm{vec}(\hat{\pi}) \end{bmatrix} \Bigg| \, Z \sim N\left(\begin{bmatrix} \gamma \\ \mathrm{vec}(\pi) \end{bmatrix}, \Sigma \otimes (Z'Z)^{-1} \right)$$

by Proposition 7.3. Letting $\tilde{\pi} = (Z'Z)^{1/2}\hat{\pi}$ and $\tilde{\gamma} = (Z'Z)^{1/2}\hat{\gamma}$ proves the result. ■

Remark 9.4. If $k = \ell = 1$, $\hat{\beta} \mid Z$ reduces to a ratio of two normals, which is known to have no moment.

9.A.2. Standard errors and inference. Heteroskedasticity is very well considered present in virtually any economic dataset. Therefore, it is always advisable to use the general formula to compute the standard error, which is valid regardless of the presence of heteroskedasticity.

Theorem 9.5 (Heteroskedasticity-robust standard error)**.** *Under Assumption 9.1 and* $\mathbb{E}[\|x_i\|^2\|z_i\|^2] < \infty$*,*

$$\left(\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 z_i z_i' \right) \xrightarrow{p} \mathbb{E}[\varepsilon_i^2 z_i z_i'].$$

With Assumption 9.2, therefore,

$$\left(\frac{1}{n} \sum_{i=1}^{n} \hat{x}_i \hat{x}_i' \right)^{-1} \hat{\pi}' \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 z_i z_i' \right) \hat{\pi} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{x}_i \hat{x}_i' \right)^{-1}$$

converges in probability to the asymptotic variance of $\hat{\beta}$ *in Theorem 9.1.*

PROOF. Note that $\frac{1}{n}\hat{X}'\hat{X} = \frac{1}{n}X'P_Z P_Z X = \frac{1}{n}X'Z(Z'Z)^{-1}Z'X = \hat{\pi}'(\frac{1}{n}Z'Z)\hat{\pi} \to^p \pi'\mathbb{E}[zz']\pi$ and $\hat{\pi} \to^p \pi$. Now it suffices to show $\frac{1}{n}\sum \hat{\varepsilon}^2 zz' \to^p \mathbb{E}[\varepsilon^2 zz']$. Since $\varepsilon - \hat{\varepsilon} = x'(\hat{\beta} - \beta)$, $\hat{\varepsilon}^2 zz' = \varepsilon^2 zz' - 2x'(\hat{\beta} - \beta)\varepsilon zz' + [x'(\hat{\beta} - \beta)]^2 zz'$. First, $\frac{1}{n}\sum \varepsilon^2 zz' \to^p \mathbb{E}[\varepsilon^2 zz']$ by the LLN. Second, using $|x'(\hat{\beta} - \beta)| \le \|x\|\|\hat{\beta} - \beta\|$, $\|\frac{1}{n}\sum x'(\hat{\beta} - \beta)\varepsilon zz'\| \lesssim \|\hat{\beta} - \beta\| \cdot \frac{1}{n}\sum |\varepsilon|\|x\|\|z\|^2 = O_P(\frac{1}{\sqrt{n}})$ since $\mathbb{E}[|\varepsilon|\|x\|\|z\|^2] \le \sqrt{\mathbb{E}[\varepsilon^2\|z\|^2]\mathbb{E}[\|x\|^2\|z\|^2]} < \infty$ by the Cauchy–Schwarz inequality. Third, $\|\frac{1}{n}\sum[x'(\hat{\beta} - \beta)]^2 zz'\| \lesssim \|\hat{\beta} - \beta\|^2 \cdot \frac{1}{n}\sum \|x\|^2\|z\|^2 = O_P(\frac{1}{n})$. This completes the proof. ■

Meanwhile, most statistical software uses as default the following standard error that is valid only under homoskedasticity. Despite Proposition 9.3, this time, there is no exact finite-sample testing we can draw from this approach. The latter property is not good enough to justify its use, however, as the assumptions have no hope to be satisfied in economic applications.

Proposition 9.6 (Default standard error)**.** *Under Assumption 9.1,*[9]

$$\left(\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \right) \left(\frac{1}{n} \sum_{i=1}^{n} \hat{x}_i \hat{x}_i' \right)^{-1} \xrightarrow{p} \mathbb{E}[\varepsilon_i^2](\mathbb{E}[x_i z_i']\mathbb{E}[z_i z_i']^{-1}\mathbb{E}[z_i x_i'])^{-1}.$$

[9]For this to be a valid asymptotic variance of $\hat{\beta}$, we need Assumptions 9.2 and 9.3.

Exercise 9.2. Prove Proposition 9.6.

The following variant of Bessel's correction is widely used in statistical software. Again, justification is only for homoskedastic models.

Proposition 9.7 (Small-sample correction)**.** *Suppose that $\hat{X}'\hat{X}$ is invertible. If $\mathbb{E}[\varepsilon_i^2 \mid \hat{X}] = \sigma^2$ for some constant σ^2, then*

$$\mathbb{E}\left[\frac{1}{n-k}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \;\middle|\; \hat{X}\right] = \sigma^2.$$

If $\mathbb{E}[\varepsilon_i^2 \mid Z, \hat{X}] = \sigma^2$, then

$$\sigma^2 A \le \mathbb{E}\left[\frac{1}{n-k}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 z_{ij}z_{ir} \;\middle|\; Z, \hat{X}\right] \le \sigma^2 B,$$

where A is the average of $n-k$ smallest elements of $\{z_{ij}z_{ir}\}_{i=1}^n$ and B of $n-k$ largest elements.

Exercise 9.3. Prove Proposition 9.7 analogously to Proposition 7.7.

Remark 9.5. The difference of the robust variance and the homoskedastic variance is $(\pi'\mathbb{E}[zz']\pi)^{-1}\pi'\operatorname{Cov}(\varepsilon^2, zz')\pi(\pi'\mathbb{E}[zz']\pi)^{-1}$. Therefore, if $\operatorname{Cov}(\varepsilon^2, zz') < 0$, the robust variance is smaller.

Given these, we can test hypotheses regarding β. Let

$$\hat{V} := \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{x}_i'\right)^{-1}\hat{\pi}'\left(\frac{1}{n-k}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 z_i z_i'\right)\hat{\pi}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{x}_i'\right)^{-1},$$

$$\hat{V}_0 := \frac{1}{n}\left(\frac{1}{n-k}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{x}_i'\right)^{-1}.$$

Theorem 9.8 (Asymptotic testing)**.** *For a full row-rank matrix R with rank r, let*

$$t_j := \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{jj}}}, \quad F_R := \frac{(R\hat{\beta} - R\beta)'(R\hat{V}R')^{-1}(R\hat{\beta} - R\beta)}{r}.$$

Under Assumptions 9.1 and 9.2 and $\mathbb{E}[\|x_i\|^2\|z_i\|^2] < \infty$, $t_j \rightsquigarrow N(0,1)$ and $rF_R \rightsquigarrow \chi^2(r)$.

Proof. It follows from Theorems 9.1 and 9.5. ■

We can substitute $\hat{V}$ with $\hat{V}_0$ to replace the assumption $\mathbb{E}[\|x\|^2\|z\|^2] < \infty$ with Assumption 9.3. Instead of a normal or chi-square, major statistical software uses critical values from a t- or F-distribution; however, there is no known theorem that justifies this practice.

Bibliography

[AAD+11] A. Abdulkadiroğlu, J. D. Angrist, S. M. Dynarski, T. J. Kane, and P. A. Pathak, *Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots*, Quarterly Journal of Economics **126** (2011), no. 2, 699–748.

[Ace09] D. Acemoglu, *Introduction to Modern Economic Growth*, Princeton University Press, Princeton, 2009.

[AJK+16] D. Acemoglu, S. Johnson, A. Kermani, J. Kwak, and T. Mitton, *The Value of Connections in Turbulent Times: Evidence from the United States*, Journal of Financial Economics **121** (2016), no. 2, 368–391.

[AL99] J. D. Angrist and V. Lavy, *Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement*, Quarterly Journal of Economics **114** (1999), no. 2, 533–575.

[Ang90] J. D. Angrist, *Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records*, American Economic Review **80** (1990), no. 3, 313–336.

[AP09] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton, 2009.

[Ayr05] I. Ayres, *Three Tests for Measuring Unjustified Disparate Impacts in Organ Transplantation: The Problem of "Included Variable" Bias*, Perspectives in Biology and Medicine **48** (2005), no. 1, S68–S87.

[BAL85] M. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge and London, 1985.

[Bas97] S. Basu, *The Conservatism Principle and the Asymmetric Timeliness of Earnings*, Journal of Accounting and Economics **24** (1997), no. 1, 3–37.

[BC11] A. Belloni and V. Chernozhukov, *High Dimensional Sparse Econometric Models: An Introduction*, Inverse Problems and High-Dimensional Estimation (P. Alquier, E. Gautier, and G. Stoltz, eds.), Springer-Verlag, Heidelberg, 2011, pp. 121–156.

[BDG+15] A. Banerjee, E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry, *A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries*, Science **348** (2015), no. 6236, 1260799.

[BDS+06] H. Benson, J. A. Dusek, J. B. Sherwood, P. Lam, C. F. Bethea, W. Carpenter, S. Levitsky, P. C. Hill, D. W. Clem, Jr., M. K. Jain, D. Drumel, S. L. Kopecky, P. S. Mueller, D. Marek, S. Rollins, and P. L. Hibberd, *Study of the Therapeutic Effects of Intercessory Prayer (STEP) in Cardiac Bypass Patients: A Multicenter Randomized Trial of Uncertainty and Certainty of Receiving Intercessory Prayer*, American Heart Journal **151** (2006), no. 4, 934–942.

[BHO75] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, *Sex Bias in Graduate Admissions: Data from Berkeley*, Science **187** (1975), no. 4175, 398–404.

[BLSZ16] A. Brodeur, M. Lé, M. Sangnier, and Y. Zylberberg, *Star Wars: The Empirics Strike Back*, American Economic Journal: Applied Economics **8** (2016), no. 1, 1–32.

[BNT15] T. Blake, C. Nosko, and S. Tadelis, *Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment*, Econometrica **83** (2015), no. 1, 155–174.

[Bre79] T. S. Breusch, *Conflict Among Criteria for Testing Hypotheses: Extensions and Comments*, Econometrica **47** (1979), no. 1, 203–207.

[Car95] D. Card, *Using Geographic Variation in College Proximity to Estimate the Return to Schooling*, Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp (L. N. Christofides, E. K. Grant, and R. Swidinsky, eds.), University of Toronto Press, Toronto, 1995, pp. 201–222.

[CB02] G. Casella and R. L. Berger, *Statistical Inference*, second ed., Cengage Learning, 2002.

[CP05] J. D. Cummins and R. D. Phillips, *Estimating the Cost of Equity Capital for Property-Liability Insurers*, Journal of Risk and Insurance **72** (2005), no. 3, 441–478.

[CR09] I. D. Coope and P. F. Renaud, *Trace Inequalities with Applications to Orthogonal Regression and Matrix Nearness Problems*, Journal of Inequalities in Pure and Applied Mathematics **10** (2009), no. 4, article 92, 7 pp.

[Dav21] J. Davidson, *Stochastic Limit Theory: An Introduction for Econometricians*, second ed., Oxford University Press, Oxford, 2021.

[DE01] H. A. David and A. W. F. Edwards, *Annotated Readings in the History of Statistics*, Springer, New York, 2001.

[DK07] S. DellaVigna and E. Kaplan, *The Fox News Effect: Media Bias and Voting*, Quarterly Journal of Economics **122** (2007), no. 3, 1187–1234.

[DS03] E. Duflo and E. Saez, *The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment*, Quarterly Journal of Economics **118** (2003), no. 3, 815–842.

[Dud14] R. M. Dudley, *Uniform Central Limit Theorems*, second ed., Cambridge University Press, Cambridge, 2014.

[EKM97] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Heidelberg, 1997.

[EKW22] G. Elliott, N. Kudrin, and K. Wüthrich, *Detecting p-Hacking*, Econometrica **90** (2022), no. 2, 887–906.

[FPP07] D. Freedman, R. Pisani, and R. Purves, *Statistics*, fourth ed., W.W. Norton & Company, New York and London, 2007.

[FW15] P. C. Flore and J. M. Wicherts, *Does Stereotype Threat Influence Performance of Girls in Stereotyped Domains? A Meta-Analysis*, Journal of School Psychology **53** (2015), no. 1, 25–44.

[GM05] T. Groseclose and J. Milyo, *A Measure of Media Bias*, Quarterly Journal of Economics **120** (2005), no. 4, 1191–1237.

[GPV00] E. Guerre, I. Perrigne, and Q. Vuong, *Optimal Nonparametric Estimation of First-Price Auctions*, Econometrica **68** (2000), no. 3, 525–574.

[Gre18] W. H. Greene, *Econometric Analysis*, eighth ed., Pearson Education, 2018.

[Han22] B. E. Hansen, *A Modern Gauss–Markov Theorem*, Econometrica **90** (2022), no. 3, 1283–1294.

[Hay00] F. Hayashi, *Econometrics*, Princeton University Press, Princeton, 2000.

[HIR03] K. Hirano, G. W. Imbens, and G. Ridder, *Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score*, Econometrica **71** (2003), no. 4, 1161–1189.

[HL07] J. Hasanhodzic and A. W. Lo, *Can Hedge-Fund Returns Be Replicated?: The Linear Case*, Journal of Investment Management **5** (2007), no. 2, 5–45.

[HLS13] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, third ed., John Wiley & Sons, Hoboken, 2013.

[HR20] M. A. Hernán and J. M. Robins, *Causal Inference: What If*, Chapman & Hall/CRC, Boca Raton, 2020.

[HS15] J. J. Heckman and M. Sattinger, *Introduction to the Distribution of Earnings and of Individual Output, by A.D. Roy*, Economic Journal **125** (2015), no. 583, 378–402.

[HS16] E. P. Herbst and F. Schorfheide, *Bayesian Estimation of DSGE Models*, Princeton University Press, Princeton and Oxford, 2016.

[HSA72] G. H. Haines, Jr., L. S. Simon, and M. Alexis, *Maximum Likelihood Estimation of Central-City Food Trading Areas*, Journal of Marketing Research **9** (1972), no. 2, 154–159.

[HSG+17] S. B. Heller, A. K. Shah, J. Guryan, J. Ludwig, S. Mullainathan, and H. A. Pollack, *Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago*, Quarterly Journal of Economics **132** (2017), no. 1, 1–54.

[IR15] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, New York, 2015.

[KCHP18] R. Koenker, V. Chernozhukov, X. He, and L. Peng (eds.), *Handbook of Quantile Regression*, CRC Press, Boca Raton, 2018.

[KMDP11] M. S. Kramer, E. E. M. Moodie, M. Dahhou, and R. W. Platt, *Breastfeeding and Infant Size: Evidence of Reverse Causality*, American Journal of Epidemiology **173** (2011), no. 9, 978–983.

[Koe05] R. Koenker, *Quantile Regression*, Cambridge University Press, Cambridge, 2005.

[KS93] R. W. Klein and R. H. Spady, *An Efficient Semiparametric Estimator for Binary Response Models*, Econometrica **61** (1993), no. 2, 387–421.

[LC98] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, second ed., Springer-Verlag, New York, 1998.

[LMW19] C. Lian, Y. Ma, and C. Wang, *Low Interest Rates and Risk-Taking: Evidence from Individual Investment Decisions*, Review of Financial Studies **32** (2019), no. 6, 2107–2148.

[LR05] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, third ed., Springer, New York, 2005.

[LW23] L. Lei and J. Wooldridge, *What Estimators Are Unbiased For Linear Models?*, arXiv:2212.14185, 2023.

[McC85] J. H. McCulloch, *On Heteros*edasticity*, Econometrica **53** (1985), no. 2, 483.

[Mee90] P. E. Meehl, *Why Summaries of Research on Psychological Theories are Often Uninterpretable*, Psychological Reports **66** (1990), no. 1, 195–244.

[MM15] F. Matějka and A. McKay, *Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model*, American Economic Review **105** (2015), no. 1, 272–298.

[MS83] W. Mellow and H. Sider, *Accuracy of Response in Labor Market Surveys: Evidence and Implications*, Journal of Labor Economics **1** (1983), no. 4, 331–344.

[MW14] S. L. Morgan and C. Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, second ed., Cambridge University Press, New York, 2014.

[NM18] T. T. Nagle and G. Müller, *The Strategy and Tactics of Pricing: A Guide to Growing More Profitability*, sixth ed., Routledge, New York and London, 2018.

[Ohl80] J. A. Ohlson, *Financial Ratios and the Probabilistic Prediction of Bankruptcy*, Journal of Accounting Research **18** (1980), no. 1, 109–131.

[Pea09] J. Pearl, *Causality: Models, Reasoning, and Inference*, second ed., Cambridge University Press, New York, 2009.

[Phi09] P. C. B. Phillips, *Exact Distribution Theory in Structural Estimation with an Identity*, Econometric Theory **25** (2009), no. 4, 958–984.

[Roy51] A. D. Roy, *Some Thoughts on the Distribution of Earnings*, Oxford Economic Papers **3** (1951), no. 2.

[RW17] J. P. Romano and M. Wolf, *Resurrecting Weighted Least Squares*, Journal of Econometrics **197** (2017), no. 1, 1–19.

[SS93] P. K. Sen and J. M. Singer, *Large Sample Methods in Statistics: An Introduction with Applications*, Springer, New York, 1993.

[SS08] R. Stinebrickner and T. R. Stinebrickner, *The Causal Effect of Studying on Academic Performance*, B.E. Journal of Economic Analysis & Policy **8** (2008), no. 1, 1–55.

[SSQ99] S. J. Spencer, C. M. Steele, and D. M. Quinn, *Stereotype Threat and Women's Math Performance*, Journal of Experimental Social Psychology **35** (1999), no. 1, 4–28.

[ST03] J. H. Stock and F. Trebbi, *Who Invented Instrumental Variable Regression?*, Journal of Economic Perspectives **17** (2003), no. 3, 177–194.

[Tri99] G. Tripathi, *A Matrix Extension of the Cauchy–Schwarz Inequality*, Economics Letters **63** (1999), no. 1, 1–3.

[TvR04] K. Talluri and G. van Ryzin, *Revenue Management Under a General Discrete Choice Model of Consumer Behavior*, Management Science **50** (2004), no. 1, 15–33.

[vdV98] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge, 1998.

[Was04] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer, New York, 2004.

[WF94] G. J. Werden and L. M. Froeb, *The Effects of Mergers in Differentiated Products Industries: Logit Demand and Merger Policy*, Journal of Law, Economics, and Organization **10** (1994), no. 2, 407–426.

[Whi82] H. White, *Maximum Likelihood Estimation of Misspecified Models*, Econometrica **50** (1982), no. 1, 1–25.

[Woo10] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, second ed., MIT Press, Cambridge, 2010.

[YFK07] Z. Yang, K.-T. Fang, and S. Kotz, *On the Student's t-Distribution and the t-Statistic*, Journal of Multivariate Analysis **98** (2007), no. 6, 1293–1304.