PhD Core Econometrics II

First edition

David M. Kaplan

First edition, Spring 2023;
Updated January 19, 2023

However, the second time round, she came upon a low curtain she had not noticed before, and behind it was a little door about fifteen inches high: she tried the little golden key in the lock, and to her great delight it fitted! Alice opened the door and found that it led into a small passage, not much larger than a rat-hole: she knelt down and looked along the passage into the loveliest garden you ever saw. How she longed to get out of that dark hall, and wander about among those beds of bright flowers and those cool fountains, but she could not even get her head through the doorway. . .

Lewis Carroll, *Alice's Adventures in Wonderland*

Brief Contents

Contents vii

Preface xi

Textbook Learning Objectives xiii

Notation 1

I Foundations 5

Introduction 7

1 Stata 9

2 Logic 13

3 The Big Picture 17

4 Identification by Independence 45

5 Identification by Conditional Independence 55

6 OVB and Proxy Variables 59

Exercises 63

II Instrumental Variables 71

7 Instrumental Variables 73

8 Generalized Method of Moments 75

Exercises 77

III Panel Data 85

9 Difference-in-Differences 87

Exercises 91

IV Probit 97

Introduction 99

10 Binary Response Models 101

Exercises 103

Bibliography 107

Index 109

Contents

Contents vii

Preface xi

Textbook Learning Objectives xiii

Notation 1

I Foundations 5

Introduction 7

1 Stata 9

1.1 Access . . . 9
1.2 Pros and Cons . . . 10
1.3 General Setup . . . 10
1.4 Administrative Commands . . . 10
1.5 Data Manipulation Commands . . . 10
1.6 Data Analysis . . . 11

2 Logic 13

2.1 Terminology . . . 13
2.2 Assumptions . . . 15
2.3 Theorems . . . 15

3 The Big Picture 17

3.1 Description, Prediction, and Causality . . . 17
3.2 Population and Sample . . . 18
3.2.1 Population Types . . . 18

3.2.2 Before and After Sampling: Two Perspectives 21
3.2.3 Sampling Types 22
3.3 Frequentist and Bayesian 25
3.3.1 Very Brief Overview: Bayesian Approach 25
3.3.2 Very Brief Overview: Frequentist Approach 26
3.3.3 Bayesian and Frequentist Differences 26
3.4 Identification, Estimation, and Inference 28
3.5 General Equilibrium and Partial Equilibrium 29
3.6 Structural and Reduced-Form Approaches 30
3.7 Linear Regression 32
3.7.1 Linear Projection 32
3.7.2 Conditional Mean Function 35
3.7.3 Causal Interpretation 36
3.8 Economic Significance 36
3.8.1 Basic Idea 36
3.8.2 Units of Measure 37
3.8.3 Log Models 37
3.9 Quantifying Uncertainty 37
3.10 Quantifying Accuracy of an Estimator 40
3.10.1 Bias 40
3.10.2 Mean Squared Error 42
3.10.3 Consistency and Asymptotic MSE 43

4 Identification by Independence 45

4.1 Average Treatment Effect 45
4.1.1 Potential Outcomes 45
4.1.2 Treatment Effects 46
4.1.3 Average Treatment Effect 47
4.1.4 ATE Identification 48
4.1.5 SUTVA Violations 50
4.2 Linear Structural Model 51
4.2.1 Fixed Coefficients 51
4.2.2 Random Coefficients 51
4.3 Nonseparable Structural Model 52

5 Identification by Conditional Independence 55

5.1 Conditional Average Treatment Effect 55
5.2 Linear Structural Model 58
5.3 Nonseparable Structural Model 58

6 OVB and Proxy Variables 59

6.1 Omitted Variable Bias . . . 59
6.2 Proxy Variables . . . 60
6.2.1 Perfect Proxy . . . 60
6.2.2 Imperfect Proxy . . . 61

Exercises **63**

II Instrumental Variables **71**

7 Instrumental Variables **73**
7.1 XXX . . . 73

8 Generalized Method of Moments **75**
8.1 XXX . . . 75

Exercises **77**

III Panel Data **85**

9 Difference-in-Differences **87**
9.1 Introduction . . . 87
9.A Technical Details: XXX . . . 89

Exercises **91**

IV Probit **97**

Introduction **99**

10 Binary Response Models **101**
10.1 XXX . . . 101

Exercises **103**

Bibliography **107**

Index **109**

Preface

This text was prepared for the 15-week 2nd-semester core PhD econometrics course at the University of Missouri. The main focus is identification, from perspectives of both structural models and potential outcomes, using conditional independence, instrumental variables (IV), or panel data. Additionally, the generalized method of moments (GMM) is discussed (after the special case of IV), as well as maximum likelihood. Probit/logit models are also presented, which allows introduction of important concepts for any nonlinear models.

The assumed background is the first-semester core PhD econometrics at the University of Missouri, which uses (roughly) the first nine chapters of Hansen (2020a) and related material from Hansen (2020b).

As with my *Introductory Econometrics* (Kaplan, 2022) and *Distributional and Nonparametric Econometrics* (Kaplan, 2021), this text's source files are freely available. Instructors may modify them as desired, or copy and paste LaTeX code into their own lecture notes, with usage subject to the Creative Commons license linked on the copyright page. I wrote the text in Overleaf, an online (free) LaTeX environment that includes knitr support. You may see, copy, and download the entire project from my website.[1]

Another unusual feature is the prevalence of in-class discussion questions. I find these very helpful (for more actively engaging students, for gauging how students are tracking, and for breaking up my lecturing), and students seem to appreciate them, too.

Thanks to everyone for their help and support: my past econometrics instructors, my colleagues and collaborators, my students, and my family.

David M. Kaplan
Spring, 2023
Columbia, Missouri, USA

[1] `https://kaplandm.github.io/teach.html`

Textbook Learning Objectives

For good reason, it has become standard practice to list learning objectives for a course as well as each unit within the course. Below are the learning objectives corresponding to this text overall. In the future, each chapter will additionally list more specific learning objectives that map to one or more of these overall objectives. I hope you find these helpful guidance, whether you are a solo learner, a class instructor, or a class student.

The textbook learning objectives (TLOs) are the following.

1. Define terms and concepts, both mathematically and intuitively.

2. Develop intuition for fundamental concepts to enable you to understand econometrics papers/books that you need to read later for your own research.

3. Describe various econometric methods both mathematically and intuitively, including their objects of interest and assumptions, and the logical relationship between the assumptions and corresponding theorems and properties.

4. For a given economic question, dataset, and econometric method, judge whether the method is appropriate and assess the economic significance and statistical significance of the results.

5. Use Stata to manipulate and analyze data, interpreting results both economically and statistically.

Notation

Variables

Usually, uppercase denotes random variables, whereas lowercase denotes fixed values. The primary exception is for certain counting variables, where uppercase indicates the maximum value and lowercase indicates a general value; e.g., time period t can be $1, 2, 3, \ldots, T$, or regressor k out of K total regressors. Scalar, (column) vector, and matrix variables are typeset differently. For example, an n-by-k random matrix with scalar (random variable) entries X_{ij} (row i, column j) is

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

and a k-dimensional non-random vector is

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{pmatrix}$$

Unless otherwise specified, vectors are column vectors. The transpose of a column vector is a row vector. For example, using the $\boldsymbol{z}$ defined above,

$$\boldsymbol{z}' = (z_1, z_2, \ldots, z_k)$$

Note: displayed math like above should always have appropriate punction (comma, period) at the end! …unless you are defining notation and worry about confusing people.

Greek letters like $\boldsymbol{\beta}$ and θ generally denote fixed population parameters.

I sometimes make exceptions to match convention. For example, ϵ is a Greek letter but is conventionally used for a regression error term or white noise.

Estimators usually have a "hat" on them. Since estimators are computed from data, they are random from the frequentist perspective. Thus, even if θ is a non-random population parameter, $\hat{\theta}$ is a random variable.

I try to put "hats" on other quantities computed from the sample, too. For example, a t-statistic would be $\hat{t}$ (a random variable computed from the sample) instead of just t (which looks like a non-random scalar). Or, a J-statistic would be $\hat{J}$, even though J is already uppercase, to emphasize that it is computed from data (rather than data itself).

Besides hats, tildes and bars may indicate estimators of parameters, and bars indicate sample averages. For example, there may be multiple alternatives for estimating θ: $\hat{\theta}$, $\tilde{\theta}$, and $\bar{\theta}$. The sample average of $Y_1, \ldots, Y_n$ is $\bar{Y}$.

Estimators and other **statistics** (i.e., things computed from data) may sometimes have a subscript with the sample size n to remind us of the asymptotic perspective of a sequence (indexed by n) of random variables. For example, with n denoting sample size, $\hat{\theta}_n$, $\hat{t}_n$, and $\bar{Y}_n$.

The following is a summary.

y	scalar fixed (non-random) value
Y	scalar random variable
θ	scalar non-random value
$\hat{\theta}$	scalar random variable
$\boldsymbol{x}$	non-random column vector
$\boldsymbol{x}'$	transpose of $\boldsymbol{x}$
$\boldsymbol{X}$	random column vector
$\boldsymbol{\beta}$	non-random column vector
$\hat{\boldsymbol{\beta}}$	random column vector
$\underline{\boldsymbol{w}}$	non-random matrix
$\underline{\boldsymbol{w}}'$	transpose of $\underline{\boldsymbol{w}}$
$\underline{\boldsymbol{W}}$	random matrix
$\underline{\boldsymbol{\Omega}}$	non-random matrix
$\hat{\underline{\boldsymbol{\Omega}}}$	random matrix

Symbols

In addition to the following symbols, vocabulary words and abbreviations (like "quantile" or "IVQR") can be looked up in the Index in the very back of the text.

$\Longrightarrow$	implies; see Chapter 2
$\Longleftarrow$	is implied by; see Chapter 2
$\Longleftrightarrow$	if and only if; see Chapter 2
$\lim_{n\to\infty}$	limit
$\operatorname{plim}_{n\to\infty}$	probability limit
$\to$	converges to (deterministic)

$\xrightarrow{p}$	converges in probability to; see Hansen (2020b, §7.3)
$\xrightarrow{\text{a.s.}}$	converges almost surely to; see Hansen (2020b, §7.14)
$\xrightarrow{d}$	converges in distribution to; see Hansen (2020b, §8.2)
$\rightsquigarrow$	converges weakly to
$\equiv$	is defined as
$\approx$	approximately equals
$\doteq$	equals when ignoring smaller-order terms
$\sim$	is distributed as
$\overset{a}{\sim}$	is distributed approximately (or asymptotically) as
$X \perp\!\!\!\perp Y$	X and Y are statistically independent
$\mathrm{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathrm{N}(0, 1)$	standard normal distribution
$\Phi(\cdot)$	cumulative distribution function (CDF) of $\mathrm{N}(0, 1)$
$\phi(\cdot)$	probability density function (PDF) of $\mathrm{N}(0, 1)$
$F_Y(\cdot)$	cumulative distribution function (CDF) of Y
$Q_Y(\cdot)$	quantile function of Y
$f_Y(\cdot)$	probability density function (PDF) of Y (or PMF if discrete)
$\mathbb{1}\{\cdot\}$	indicator function: $\mathbb{1}\{A\} = 1$ if event A occurs, else $\mathbb{1}\{A\} = 0$
$\mathrm{P}(A)$	probability of event A
$\mathrm{P}(A \mid B)$	conditional probability of A given B
$\mathrm{E}(Y)$	expected value of Y
$\widehat{\mathrm{E}}(Y)$	expectation for sample distribution; same as $\frac{1}{n}\sum_{i=1}^{n} Y_i$
$\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$	CEF (function of $\boldsymbol{x}$); see Hansen (2020a, §2.5)
$\mathrm{E}(Y \mid \boldsymbol{X})$	expected value of Y given $\boldsymbol{X}$; this is a random variable
$\mathrm{Var}(Y)$	variance of Y
$\mathrm{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x})$	conditional variance (a non-random value)
$\mathrm{Var}(Y \mid \boldsymbol{X})$	conditional variance (a random variable)
$\mathrm{Cov}(Y, X)$	covariance
$\mathrm{Corr}(Y, X)$	correlation
$b \in \{a, b, c\}$	b is in the set containing a, b, and c
$\mathcal{S}_1 \cup \mathcal{S}_2$	the union of sets $\mathcal{S}_1$ and $\mathcal{S}_2$
$\bigcup_{j=1}^{J} \mathcal{S}_j$	the union of $\mathcal{S}_1, \ldots, \mathcal{S}_J$
$\mathcal{S}_1 \cap \mathcal{S}_2$	the intersection of sets $\mathcal{S}_1$ and $\mathcal{S}_2$
$\bigcap_{j=1}^{J} \mathcal{S}_j$	the intersection of $\mathcal{S}_1, \ldots, \mathcal{S}_J$
$\mathbb{N}$	the set of natural numbers, $\{1, 2, 3, \ldots\}$
$\mathbb{R}$	the set of real numbers (which excludes $\pm\infty$)
$\mathbb{R}_{\geq 0}$	the non-negative real numbers
$\mathbb{R}_{>0}$	the strictly positive real numbers
$\bar{\mathbb{R}}$	the extended real numbers, $\mathbb{R} \cup \{-\infty, \infty\}$

$\mathbb{R}^k$	k-dimensional Euclidean space
$\mathbb{Z}$	the set of integers, $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{Z}_{\geq 0}$, $\mathbb{Z}_{>0}$	analogous to $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$
$\mathrm{SE}(\hat{\theta})$	standard error of estimator $\hat{\theta}$
$\arg\min_g f(g)$	the value of g that minimizes $f(g)$
$\underset{\sim}{\boldsymbol{I}}_k$	$k \times k$ identity matrix (ones on main diagonal, zeros elsewhere)
$\lVert\cdot\rVert$	norm (Euclidean unless otherwise defined)
$\mathrm{tr}(\underset{\sim}{\boldsymbol{v}})$	trace of matrix $\underset{\sim}{\boldsymbol{v}}$
$\underset{\sim}{\boldsymbol{v}}'$	transpose of matrix $\underset{\sim}{\boldsymbol{v}}$
$\underset{\sim}{\boldsymbol{v}}^{-1}$	inverse of matrix $\underset{\sim}{\boldsymbol{v}}$
$\underset{\sim}{\boldsymbol{v}} > 0$	matrix $\underset{\sim}{\boldsymbol{v}}$ is positive definite
$\underset{\sim}{\boldsymbol{v}} \geq 0$	matrix $\underset{\sim}{\boldsymbol{v}}$ is positive semi-definite

Part I

Foundations

Introduction

This part may be largely review, but it is helpful to have a deeper understanding of "basic" ideas before adding complexity. Eventually the focus narrows to identification of causal effects, specifically how in linear regression "control variables" can help reduce omitted variable bias but usually do not eliminate it.

Chapter 1

Stata

Unit learning objectives for this chapter

1.1. Access Stata and code basic commands for data manipulation and analysis. [TLO 5]

This chapter provides a brief overview of Stata, which you will use for the end-of-part exercises in this book.

> Optional resources for this chapter
>
> - Many user-contributed Stata commands can be installed from SSC, including `bcuse` (Baum, 2012), `ivreg2` (Baum, Schaffer, and Stillman, 2002), and `ranktest` (Kleibergen, Schaffer, and Windmeijer, 2007), which are used in this class.
> - UCLA resources: https://stats.idre.ucla.edu/stata

1.1 Access

As a student at Mizzou, you can use Software Anywhere for free, even from home.[1] It currently has Stata version 15 (StataCorp, 2017), which is a few versions old but sufficient for this class.

The on-campus computing sites also provide a variety of statistical software. You can check which computing sites/labs have your favorite software on the Computing Sites Software web page.[2]

[1] https://doit.missouri.edu/services/software/software-anywhere/

[2] https://doit.missouri.edu/services/computing-sites/sites-software/

1.2 Pros and Cons

As with econometric paradigms, different statistical software packages have complementary strengths; none is "best" for every case. Stata is commonly used by eocnomists, especially in applied microeconomics. Below are some general strengths and weaknesses.
Strengths:

1. Very intuitive and simple; easy to do most common tasks.
2. Popular among applied economists $\Longrightarrow$ lots of support, data often available in Stata format, used in jobs, etc.
3. I think the help files within Stata are very helpful (once you know the basic structure and syntax).

Weaknesses:

1. Not as many fancy functions as R, although econometricians are getting better about providing code in Stata (e.g., lots of the new RD methods).
2. Not as easy to code your own functions (vs. R, based on my experiences doing both).
3. Can only have one dataset in memory at a time.
4. Can be slower, but depends on Stata version (some support parallel processing) and the particular computational task.

1.3 General Setup

XXX GUI layout/panes/windows

XXX ssc / ado files

XXX do-files (incl replicability)

1.4 Administrative Commands

XXX comments

XXX version/which

XXX set more off (etc.)

XXX cd/pwd

XXX logs (.log)

XXX clear/load/save data, and insheet/etc.

XXX describe / list in 1/5 / browse (/edit)

XXX conclusion: basic template...

1.5 Data Manipulation Commands

XXX keep, drop, order

XXX keep if / drop if

XXX sort/gsort

XXX reshape (wide/long)
XXX append
XXX merge 1:1/m:1/1:m, _merge
XXX generate/replace (arithmetic, etc.)
XXX collapse

1.6 Data Analysis

XXX summarize
XXX tabulate
XXX regress, etc.....more later
XXX graphs (histogram, scatter)

Chapter 2

Logic

Unit learning objectives for this chapter

2.1. Define and apply basic logic terms and relationships [TLO 1]

Some basic logic is useful for understanding certain parts of econometrics. First, logic is useful for understanding the relationship among different conditions. Often these conditions are assumptions used in various theorems. Second, logic is useful for understanding what a theorem actually claims. Third, logic is helpful for interpreting results. The following may not be fully technically correct from a philosopher's perspective, e.g., perhaps I conflate logical implication with the material conditional, but it suffices for econometrics.

Optional resources for this chapter

- Section 6.1 of Kaplan (2022) is very similar, and Chapter 3 of Kaplan (2021) is identical

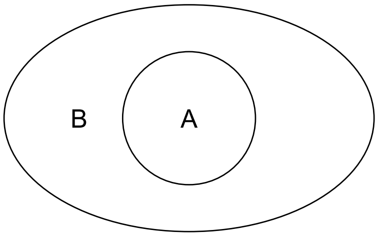
2.1 Terminology

Many words and notations can refer to the same logical relationship. Let A and B be two statements that can be either true or false. For example, maybe A is "$Y \geq 10$" and B is "$Y \geq 0$." Or, A is "this animal is a cat," and B is "this animal is a mammal." The following ways of describing the logical relationship between A and B all have the same meaning.

1. If A (is true), then B (is true)
2. $A \implies B$

3. A **implies** B
4. $B \impliedby A$
5. B is **implied by** A
6. B is true **if** A is true
7. A is true **only if** B is true
8. A is a sufficient condition (or just **sufficient**) for B
9. B is a necessary condition (or just **necessary**) for A
10. A is **stronger** than B
11. B is **weaker** than A
12. It is impossible for B to be false when A is true (but it is fine if both are true, or both are false, or A is false and B is true)
13. The truth table (T=true, F=false):

A	B	$A \implies B$
T	T	T
T	F	F
F	T	T
F	F	T

14. 

To state equivalence of A and B, opposite statements can be combined. Specifically, any of the following have the same meaning:

1. $A \iff B$ (meaning both $A \implies B$ and $A \impliedby B$)
2. A is true **if and only if** B is true (meaning A is true if B is true *and* A is true only if B is true)
3. A is necessary and sufficient for B (or equivalently, B is necessary and sufficient for A)
4. A is equivalent to B
5. It is impossible for A to be false when B is true, and impossible for A to be true when B is false.
6. The truth table (T=true, F=false):

A	B	$A \iff B$
T	T	T
T	F	F
F	T	F
F	F	T

Variations of $A \implies B$ have the following names. Read $\neg A$ as "not A": $\neg A$ is false when A is true, and $\neg A$ is true when A is false.

- $\neg A \implies \neg B$ is the **inverse** of $A \implies B$.
- $B \implies A$ is the **converse** of $A \implies B$.
- $\neg B \implies \neg A$ is the **contrapositive** of $A \implies B$.

Interestingly, the statement $A \implies B$ is logically equivalent to its contrapositive. That is, statements "$A \implies B$" and "$\neg B \implies \neg A$" can be both true or both false, but it's impossible for one to be true and the other false. The statement $A \implies B$ is not logically equivalent to either its inverse or converse. (The inverse and converse are equivalent to each other: the inverse is the contrapositive of the converse.)

Discussion Question 2.1 (logic)**.** Let A be "$X \leq 0$" and let B be "$X \leq 10$."
a) Explain why $A \implies B$.
b) State the contrapositive in terms of X, and explain why it is also true.
c) State the converse in terms of X, and explain why it is not true.
d) State the inverse in terms of X, and explain why it is not true.

2.2 Assumptions

To compare assumptions, the terms "stronger" and "weaker" are most commonly used. Instead of assumption A and conclusion B, let A and B denote different assumptions. For example, let A be $\mathrm{E}(Y^4) < \infty$, and let B be $\mathrm{E}(Y^2) < \infty$. Any random variable Y with finite $\mathrm{E}(Y^4)$ also has finite $\mathrm{E}(Y^2)$, but some have finite $\mathrm{E}(Y^2)$ and infinite $\mathrm{E}(Y^4)$. Logically, $A \implies B$. Thus, people say "$\mathrm{E}(Y^4) < \infty$ is a stronger assumption than $\mathrm{E}(Y^2) < \infty$," or equivalently, "$\mathrm{E}(Y^2) < \infty$ is weaker than $\mathrm{E}(Y^4)$."

As another example, consider the linear projection and linear CEF models. Consider the linear model $Y = \beta_0 + \beta_1 X + U$. Let assumption A be $\mathrm{E}(U \mid X) = 0$, and let B be $\mathrm{E}(U) = 0$ and $\mathrm{Cov}(X, U) = 0$; i.e., A says U is a CEF error, whereas B says U is a linear projection error. Here, $A \implies B$, so A is a stronger assumption than B, and B is weaker than A. Seen another way, the linear projection model is more general than the linear CEF model: if the CEF is $\beta_0 + \beta_1 x$, then so is the linear projection, but if the linear projection is $\beta_0 + \beta_1 x$, it is still possible to have a nonlinear CEF.

All else equal, weaker assumptions are better because then the theorem applies to more settings (the results are "more general").

2.3 Theorems

Theorems all have the same logical structure: if assumption A is true, then result (conclusion) B is true. Sometimes A and B have multiple parts, like the four parts of Assumption 7.1 of Hansen (2020a, §7.1, p. 170) and the five conclusions in Theorem 7.1 of Hansen

(2020a, §7.2, p. 172), but the logical structure of a theorem is always the same. The theorem claims that if we can verify that A is true, then we know that B is also true. But what if we don't know about A, or we think it's false? Then, B could be false, or it could be true. This may be seen most readily from the picture version of the A and B relationship. We could be somewhere inside B (where B is true) but outside A (where A is false); or we could be outside both, where both are false. The theorem is not equivalent to, "If A is false, then B is false" (the "inverse"). However, it is equivalent to the **contrapositive**: "If B is false, then A is false." Again, this is probably seen most easily in the picture.

Discussion Question 2.2 (median theorem logic)**.** Consider the statement, "If sampling is iid, then the sample median consistently estimates the population median."

a) What does this tell us about consistency of the sample median when sampling is not iid?
b) What does this tell us about sampling when the sample median is not consistent?

Hint: draw a picture.

Discussion Question 2.3 (mean theorem logic)**.** Consider the statement, "If sampling is iid and the population mean is well-defined, then the sample mean consistently estimates the population mean."

a) What does this tell us about consistency of the sample mean when sampling is not iid?
b) What does this tell us about sampling when the sample mean is not consistent?

Hint: draw a picture with A1 (iid), A2 (well-defined), and B (consistency).

Discussion Question 2.4 (logic with feathers)**.** Consider two theorems. Theorem 1 says, "If X is an eagle, then it has feathers." Theorem 2 says, "If X is a bird, then it has feathers."

a) Describe each theorem logically: what's the assumption (A), what's the conclusion (B), what's the relationship?
b) State Theorem 1's contrapositive; is it true?
c) Compare: does Theorem 1 or Theorem 2 have a stronger assumption? Why?
d) Compare: which theorem is more useful? (Which applies to more situations?)

Chapter 3
The Big Picture

Unit learning objectives for this chapter

3.1. Define terms and concepts fundamental to econometrics as a whole and the portion of it on which we will focus, including the interpretation and significance of empirical results. [TLOs 1, 2, and 4]

This chapter provides a view of the wide world of econometrics, including fundamental ideas that will recur throughout the book. The section titles use the word "and" instead of "versus" to emphasize that different paradigms may be helpful in different contexts or even complement each other in the same context; it is not a fight about which is "best," because there is no universal "best." If you end up using econometrics for research, then you will (hopefully) not be using methods directly from this class but more sophisticated methods that you learn about later. I hope this chapter (and book) helps you more readily understand the new methods you encounter later.

3.1 Description, Prediction, and Causality

This section draws from Section 4.3 of Kaplan (2022).

There are three categories of questions that econometric methods can help answer, related to description, prediction, and causality. Description is essentially about features of the joint distribution of observable variables, like correlations and conditional means. Prediction is guessing an unknown value based on other observed values; the "best" guess depends on the consequences of wrong guesses, which are often used to make a decision. Causality is important for making decisions, like should our firm spend more on advertising, or should we raise or lower the minimum wage? Causality is about the effect of such policy changes on other variables.

Example 3.1 (Kaplan video)**.** Consider the relationship between an individual's employment status and mental health, specifically anxiety. A descriptive question is: what's the proportion of employed individuals who have generalized anxiety disorder (GAD), and how much higher or lower is that proportion among unemployed individuals? A predictive question is: given somebody's employment status, what's the "best" guess of their score on the GAD-7 anxiety measure? A causal question is: how does being employed (instead of unemployed) affect an individual's level of anxiety as quantified by the GAD-7?

Discussion Question 3.1 (description, prediction, causality)**.** Which type of question (description, prediction, causality) is each of the following? Explain why. Hint: there's one of each.

a) If you only know whether an individual is from Canada or the U.S., what is your best guess of their income?
b) You are currently working in the U.S. but considering moving to Canada. How will your income change if you do?
c) Which country's population has higher income: Canada or the U.S.?

3.2 Population and Sample

Generally, the population is what we want to learn about, and the sample is the data from which we can learn. There are different ways to mathematically model a population, depending on the object of interest. Often, we can learn about a feature of a population by computing the same feature of the sample. More details about how we interpret the population and sample are in Section 3.4.

In this book, the population is a represented by a joint probability distribution of random variables, and the sample is a set of n observations of values drawn from that distribution. Population features are mirrored by features of the sample. For example, population random variable Y has mean $\mathrm{E}(Y)$, and given $Y_1, Y_2, \ldots, Y_n$, the sample mean is $\widehat{\mathrm{E}}(Y) = (1/n)\sum_{i=1}^{n} Y_i$. The **sample distribution** or **empirical distribution** is a discrete probability distribution with probability $1/n$ on each value of Y_i; the sample mean is thus the mean of the empirical distribution.

3.2.1 Population Types

$\Longrightarrow$ Kaplan video: Population Types

This subsection is a shorter version of Section 2.2 of Kaplan (2022).

In this textbook, the population is modeled mathematically as a probability distribution. This is appropriate for the infinite population or superpopulation below, but not the finite population. Consequently, it is most important to distinguish between the finite population and the other two types.

Beyond our scope...

Recently, there has been some renewed interest in finite-population methods in econometrics; for example, see Abadie, Athey, Imbens, and Wooldridge (2020).

Generally, the finite population perspective cares more about the outcomes of a finite group of individuals, whereas the other two population types care more about properties of the underlying mechanisms that generated the outcomes, often called the **data-generating process** (DGP).

A **finite population** is closest to the regular English word "population," which means all the people living in some area. For example, if we are interested in the outcomes of (only) everyone in Missouri in 2023, then we have a finite population. Other examples of finite populations are (for a given time period) all employees at a particular firm, all firms in a particular industry, all students in a particular school, or all hospitals of a certain size. In a finite population, we care only about the actual outcomes, not underlying reasons; for example, maybe we want to know how many individuals in Missouri actually earned the minimum wage in January 2023, but we do not care about the determinants of wage. Hypothetically, if we could observe every single member of a finite population, then we could fully answer our question, with no uncertainty. That is, our confidence interval would just be a single point, equal to the true value.

Sometimes a finite population is so large compared to the sample size (i.e., the number of population members we observe) that an **infinite population** is a reasonable approximation. For example, if we observe only 600 individuals out of the 6+ million in Missouri, econometric results based on finite and infinite populations are practically identical. Although "infinite" sounds more complex than "finite," it is actually simpler mathematically: instead of needing to track every single member of a finite population, an infinite population is succinctly described by a probability distribution or random variable.

Besides this convenience, sometimes there is no finite population (however large) that answers your question. For example, imagine there's a new manufacturing process for carbon monoxide monitors that should sound an alarm above 50ppm. Most work properly, but some are faulty and never alarm. Specifically, this manufacturing process corresponds to some probability of producing a faulty monitor. Mathematically, the manufacturing process can be modeled as random variable W with some probability of the value "faulty." If you want to learn this probability (i.e., this property of the manufacturing process), then there is no finite number of monitors that can exactly answer your question; no finite number of realizations exactly determines $\mathrm{P}(W = \text{faulty})$. This is an infinite population question.

One variation of the infinite population is the **superpopulation** (coined by Deming and Stephan, 1941). This imagines (infinitely) many possible universes; our actual universe is just one out of infinity. Thus, even if it appears we have a finite population, we could imagine that our universe's finite population is actually a single sample from an infinite number of universes' finite populations. The term "superpopulation" essentially means "population of populations." Our universe's finite population "is only one of the many possible populations that might have resulted from the same underlying system of social and economic causes" (Deming and Stephan, 1941, p. 45). For example, imagine we want to learn the relationship between U.S. state-level unemployment rates and state minimum wage levels. It may appear we are stuck with a finite population because there are only 50 states, each of which has an observable unemployment rate and minimum wage. However, observing all 50 states still doesn't fully answer our question about the underlying mechanism that relates unemployment and minimum wage, so a finite population seems inappropriate. But we can't just manufacture new states like we can manufacture new carbon monoxide monitors, so an infinite population also seems inappropriate. The superpopulation imagines manufacturing new entire universes, each with 50 states and the same economic and legal systems.

In Sum: Population Type

Hypothetically, could a finite number of observations fully answer your question?
No $\implies$ superpopulation or infinite population, modeled as probability distribution (as in this textbook)
Yes $\implies$ finite population (use different methods unless sample is much smaller than population)

Example 3.2 (employment status)**.** Consider the employment status of individuals in Missouri. A finite population is more appropriate if you want to document the actual percentage of Missouri individuals unemployed last week. A superpopulation is more appropriate if you want to learn about the underlying mechanism that relates education and unemployment. That is, knowing each individual's employment status fully answers the first question, but not the second question.

Example 3.3 (employee productivity)**.** Consider the productivity of employees at your company (you're the CEO). If you want to know each employee's productivity over the past fiscal quarter, then a finite population is more appropriate. If you want to learn how a particular company policy affects productivity, then a superpopulation is more appropriate. That is, knowing each employee's productivity fully answers the first question, but not the second question.

Discussion Question 3.2 (student data)**.** Imagine you're a high school principal. You have data on every student, including their standardized test scores from last spring.

a) Describe a specific question for which the finite population is most appropriate, and explain why.
b) Describe a specific question for which an infinite population or superpopulation is most appropriate, and explain why.

3.2.2 Before and After Sampling: Two Perspectives

⟹ Kaplan video: "Before" and "After" Perspectives of Data

This subsection is from Section 2.1 of Kaplan (2022).

Consider a coin flip. The two possible outcomes are heads (h) and tails (t). After the flip, we observe the outcome (h or t). Before the flip, either h or t is possible, with different probabilities.

Let variable W represent the outcome. After the flip, the outcome is known: either $W = h$ or $W = t$. Before the flip, both $W = h$ and $W = t$ are possible. If the coin is "fair," then possible outcome $W = h$ has probability $1/2$, as does $W = t$.

The "after" view sees W as a **realized value** (or **realization**). It is either heads or tails. Even if the actual "value" (heads or tails) is unknown to us, there is just a single value. For example, in physics the variable c represents the speed of light in a vacuum; you may not know the value, but c represents a single value.

Instead, the "before" view sees W as a **random variable**. That is, instead of representing a single (maybe unknown) value like in algebra, W represents a set of possible values, each associated with a probability. In the coin flip example, the possible outcomes are h and t, and the associated probabilities are both 0.5.

Other terms for W include a **random draw** (or just **draw**), or more specifically a random draw (or "randomly drawn") from a particular probability distribution. Seeing the population as a probability distribution (see Section 3.2.1), we could say W is randomly sampled from its population distribution, or if there are multiple random variables $W_1, W_2, \ldots$ (e.g., multiple flips of the same coin), we could say they are randomly sampled from the population or that they collectively form a **random sample**; see Section 3.2.3 for more about sampling.

Notationally, in this textbook, random variables are usually written uppercase (like W or Y), whereas realized values are usually written lowercase (like w or y). This notation is not unique to this textbook, but beware that other books use different notation. (For more on notation, see the Notation section in the front matter before Chapter 1.)

Example 3.4 (Kaplan video)**.** Let $R = 1$ if it rains in Columbia, MO on Tuesday and $R = 0$ if not. If today is Monday, then either outcome is possible, so we have the "before" view: R is a random variable, with some probability of $R = 0$ and some probability of $R = 1$. If instead today is Wednesday, then what happened Tuesday is already determined, so we have the "after" view. If it rained, then $R = 1$; if not, $R = 0$. There is only a single value, not multiple possible values. Even if we don't know the realized value r, we know it's just a single value.

Extending the above are the **before sampling** and **after sampling** perspectives, or "before observation" and "after observation." Similar to above, "before" corresponds to random variables, whereas "after" corresponds to realized values. Before sampling a unit (person, firm, etc.) from a population, we don't know which one we'll get, so there are multiple possible values. After sampling, we can see the specific values we got.

Example 3.5 (age as random variable)**.** Imagine you plan to record the age of one person living in your city. You take a blank piece of paper on which you'll write the age. After you find a person and write their age ("after sampling"), that number can be seen as a realized value, like w. In contrast, before sampling, there are many possible numbers that could end up on your paper. It's not that peoples' ages are undetermined; they each know their own age. But before you "sample" somebody, it's undetermined whose age will end up on your paper. It could be your neighbor DeMarcus, age 88. It could be your kid's friend Lucia, age 7. It could be your colleague Xiaohong, age 35. The random variable W is like your blank paper: it has many possible values, like $W = 88$, $W = 7$, or $W = 35$.

Discussion Question 3.3 (web traffic)**.** Let $Y = 1$ if you're logged into the course website and $Y = 0$ if not.

a) From what perspective is Y a non-random value?
b) From what perspective is Y a random variable?

In Sum: Before & After

Before: multiple possible values $\Longrightarrow$ random variable
After: single observed value $\Longrightarrow$ realized value (non-random)

3.2.3 Sampling Types

$\Longrightarrow$ Kaplan video: Types of Sampling

This subsection is a shorter version of Section 3.2 of Kaplan (2022).

Properties of estimators depend on how a sample is drawn from the population. However, this book focuses mostly on identification, so generally iid sampling (see below) is assumed for simplicity. One exception is the discussion of "cluster-robust" confidence intervals when using panel data. There are also problems related to sampling like sample selection bias and missing data; for example, see Section 12.3 ("Threats to Internal Validity") of Kaplan (2022) or Chapter 21 ("Missing Data") of Kaplan (2021).

Notationally, we observe the values from n **units**, which could be individuals, firms, countries, etc. (I often refer to units as "individuals," too.) Let $i = 1$ refer to the first unit, $i = 2$ to the second, etc., up to $i = n$, where n is the **sample size**. The corresponding values are $Y_1, Y_2, \ldots, Y_n$, with Y_i more generally denoting the observation for unit i. A particular dataset may have specific values like $Y_1 = 5$, $Y_2 = 8$, etc., but to analyze (frequentist) statistical properties, each Y_i is seen as a random variable as in Section 3.2.2. You can imagine n buckets (or pieces of paper), initially empty, that will eventually contain information from n observations. The sampling procedure does not determine the specific numeric values that end up in the buckets, but it determines how the buckets get filled.

In this section, two important sampling properties are considered: "independent" and "identically distributed." If both hold, then the Y_i are called **independent and identically distributed** (iid) random variables (or "sampled iid"), and "sampling is iid." Sometimes the vague phrase **random sampling** refers to iid sampling.

Notationally, iid sampling is indicated by $\overset{iid}{\sim}$. For example, with population CDF $F_Y(\cdot)$,

$$Y_i \overset{iid}{\sim} F_Y, \quad i = 1, \ldots, n. \tag{3.1}$$

The F_Y can be replaced by another distribution function or name.

Independent

Qualitatively, in the context of sampling, **independence** (or independent sampling) means that from the "before" view, any two observations are unrelated. For example, the value of Y_2 is unrelated to Y_1: we are not any more likely to see a high Y_2 if we see a high Y_1 in the sample.

Mathematically, independence means

$$Y_i \perp\!\!\!\perp Y_k \text{ for any } i \neq k, \tag{3.2}$$

where $\perp\!\!\!\perp$ denotes statistical independence. That is, $Y_1 \perp\!\!\!\perp Y_2$, $Y_1 \perp\!\!\!\perp Y_8$, $Y_6 \perp\!\!\!\perp Y_4$, etc. For any $i \neq k$, independent sampling implies (but is not implied by), among other properties,

$$\mathrm{Cov}(Y_i, Y_k) = 0, \quad \mathrm{Var}(Y_i + Y_k) = \mathrm{Var}(Y_i) + \mathrm{Var}(Y_k), \quad \mathrm{E}(Y_i \mid Y_k) = \mathrm{E}(Y_i). \tag{3.3}$$

Example 3.6 (Kaplan video)**.** You plan to flip a coin and record $Y_1 = 1$ if heads and $Y_1 = 0$ if tails. You plan flip the same coin again and record $Y_2 = 1$ if heads and $Y_2 = 0$ if tails. These are independent: $Y_1 \perp\!\!\!\perp Y_2$. Although the probabilities are very closely related (actually identical), the realization of the first flip (heads or tails) has no relationship with the second flip. For example, even if we know the first flip is heads, this does not change the probability of heads for the second flip: $\mathrm{P}(Y_2 = 1 \mid Y_1 = 1) = \mathrm{P}(Y_2 = 1)$.

Example 3.7 (Kaplan video)**.** You plan to pick a random person in the world and record how many years of formal education they've had as Y_1. You plan to then pick another random person and record their years of education in Y_2. The way you sample Y_2 has no relation to the first sampled person or their Y_1 value, so there is independence: $Y_1 \perp\!\!\!\perp Y_2$. Among other implications, this means Y_1 and Y_2 have zero correlation (uncorrelated) and zero covariance, $\mathrm{Cov}(Y_1, Y_2) = 0$.

Identically Distributed

The **identically distributed** property means that from the "before" view, the distribution of Y_i is the same for any i. Qualitatively, all units are sampled from the same population. Mathematically, given shared population CDF $F_Y(\cdot)$, $Y_i \sim F_Y$ for all $i = 1, \ldots, n$; or without specifying F_Y explicitly, identically distributed means $Y_i \overset{d}{=} Y_k$ for any i, k. This further implies equalities like $\mathrm{E}(Y_i) = \mathrm{E}(Y_k)$ and $\mathrm{Var}(Y_i) = \mathrm{Var}(Y_k)$.

Example 3.8 (Kaplan video)**.** The Y_1 and Y_2 in Example 3.6 are identically distributed because they are from the same coin, so the probability of heads is the same each time. (Unless you cheat or flip it differently or something, but those are nuances for physics class, not econometrics.)

Discussion Question 3.4 (i/id sampling)**.** You are planning to sample values Y_1 and Y_2, but you have not yet sampled them. Each of the following four statements implies one of the four sampling properties: 1) independent, 2) not independent (i.e., dependent), 3) identically distributed, 4) not identically distributed. Which is which?

a) You are just as likely to get $Y_1 = 3$ as $Y_2 = 3$, and similarly for any other value besides 3.
b) If you get a negative Y_1, then you'll probably get a negative Y_2; but if you get a positive Y_1, then you'll probably get a positive Y_2.
c) Separately and simultaneously, you will randomly sample Y_1 while your friend samples Y_2.
d) For Y_1 you are going to get the salary of somebody with an economics degree, and Y_2 will be the salary of somebody with an art history degree.

Example 3.9 (Kaplan video)**.** Imagine randomly picking a Mizzou student ID number, then randomly picking a 2nd, then 3rd, then 4th. The corresponding Y_i are both independent and identically distributed (iid). They are independent because each ID number is randomly drawn without any consideration of how the other numbers are drawn, and without any consideration of the other observed Y_i values. They are identically distributed because each ID number is drawn from the same population (anyone who has a Mizzou student ID).

Example 3.10 (Kaplan video)**.** Each Mizzou student is classified as either a resident of Missouri ("in-state") or not ("non-resident"). Imagine buckets 1 and 2 say "in-state," while buckets 3 and 4 say "non-resident": observations Y_1 and Y_2 are from in-state students, while Y_3 and Y_4 are from non-resident students. (This is "stratified sampling": assigning buckets to different strata before sampling.) For most variables, the in-state distribution differs from the non-resident distribution, so the distribution of Y_1 and Y_2 (in-state) differs from the distribution of Y_3 and Y_4 (non-resident). That is, sampling is not identically distributed. Thus, even if the samples are all independent, sampling is not iid.

Example 3.11 (Kaplan video)**.** Imagine randomly picking a class (like introductory econometrics) at Mizzou, and filling the first two buckets (Y_1 and Y_2) with two random students from that class; then randomly picking another class, and another two students for the other buckets (Y_3 and Y_4). (This is an example of "clustered sampling," where each class is a "cluster"; this differs from "clustering" in cluster analysis.) Observations are identically distributed (because each Y_i has the same probability of getting any particular student) but probably not independent. For example, dependence may come from students in the same class being similarly affected by their shared experience. Here, buckets 1 and 2 are correlated, and 3 and 4 are correlated, but not 1 and 3, nor 2 and 4, etc. Thus, sampling is not iid.

Example 3.12 (Kaplan video)**.** Imagine randomly picking 2 Mizzou students (like with random ID numbers), then observing them this semester and next semester. For example, imagine bucket 1 contains the first student's GPA this semester, bucket 2 contains the same student's GPA next semester, and buckets 3 and 4 contain the other student's GPAs from this semester and next semester. Buckets 1 and 2 (Y_1 and Y_2) are probably both high or both low, rather than one high and one low, and similarly for buckets 3 and 4 (Y_3 and Y_4). That is, buckets 1 and 2 are correlated, and 3 and 4 are correlated. Further, observations may not even be identically distributed if fall GPA and spring GPA do not have the same distribution. Thus, sampling is not iid.

Discussion Question 3.5 (rural household sampling)**.** You want to learn about household consumption in rural Indonesia. In an area with 100 villages, you either i) pick 5 villages at random, then survey every household in each of the 5 villages; or ii) make a list of all households in all 100 villages, then randomly pick 5% of them. Explain why each approach is or isn't iid.

3.3 Frequentist and Bayesian

$\Longrightarrow$ Kaplan video: Bayesian and Frequentist Perspectives

This subsection is a shorter version of Section 3.1 of Kaplan (2022).

The **Bayesian** and **frequentist** (or **classical**) frameworks have both produced valuable econometric methods. This book uses the frequentist framework. Often the practical difference is small, although in some cases it can be large (e.g., Kaplan and Zhuo, 2021).

The goal of this section is to develop a basic understanding of both frameworks, including how sampled data is used to learn about the population, as well as how uncertainty is quantified.

3.3.1 Very Brief Overview: Bayesian Approach

The Bayesian approach models your beliefs about an unknown population value θ, like the mean $\theta = \mathrm{E}(Y)$. Your **prior** (or prior belief) is what you believe about θ before seeing the data. Your **posterior** (or posterior belief) is what you believe about θ after seeing the data. The Bayesian approach describes how to update your prior using the observed data, to get your posterior.

Mathematically, "belief" is a probability distribution. For example, let random variable B represent your belief about the population mean. If you think there's a 50% chance the mean is negative, then $\mathrm{P}(B < 0) = 50\%$. If you think there's a 1/4 probability that B is below -1, then $\mathrm{P}(B < -1) = 1/4$. (Elsewhere, you may see this written more confusingly as $\mathrm{P}(\theta < 0)$ and $\mathrm{P}(\theta < -1)$.)

For example, imagine you see a bird flying in your backyard, and you grab your binoculars to try to identify it. Let θ represent the true species, while B is your belief.

Imagine (for simplicity) you only ever see three types of bird in your backyard, all woodpeckers: downy, hairy, and red-bellied, written $\theta = d$, $\theta = h$, and $\theta = r$. Based on the location and habitat, you know hairy is somewhat less likely in general, so your prior is $\mathrm{P}(B = d) = \mathrm{P}(B = r) = 0.4$, $\mathrm{P}(B = h) = 0.2$. Looking through your binoculars (looking at the data), you're pretty sure it's not the red-bellied, but it's too far to distinguish downy from hairy, so your updated posterior belief has $\mathrm{P}(B = d) = 0.6$, $\mathrm{P}(B = h) = 0.3$, $\mathrm{P}(B = r) = 0.1$. The low probability of red-bellied comes from the data, whereas the higher probability of downy than hairy comes from your prior.

The posterior distribution is the Bayesian way of quantifying uncertainty. It is relatively intuitive, similar to how people talk about uncertainty in daily life. The posterior distribution is often summarized by a **credible interval**, i.e., a range of values that you're pretty sure (like 90% sure) contains the true θ. Or in the above example with categorical θ, the **credible set** $\{d, h\}$ has 90% posterior belief: you'd say, "I'm 90% sure it's a downy or hair woodpecker, although I think there's a 10% chance I'm wrong and it's a red-bellied woodpecker."

3.3.2 Very Brief Overview: Frequentist Approach

The core of the frequentist approach is the "before" perspective (Section 3.2.2), which can also be described in terms of **repeated sampling**. Instead of the belief probabilities of a Bayesian posterior, frequentist probabilities are from the "before" view of the dataset (and thus value of estimator and such). Equivalently, as a thought experiment, we can imagine many different random samples drawn from the same population; the "before" probabilities are then how often certain values occur in these many random datasets.

For intuition, imagine you could randomly sample 100 datasets from the same population. Then, the frequentist probability of an event is approximately how many times that event occurs among the 100 samples. For example, we could compute the sample mean $\bar{Y}$ in all 100 samples; because the datasets are all different, the sample means $\bar{Y}$ are also all different. If $\bar{Y} \leq 0$ in 50 of the 100 hypothetical samples, then $\mathrm{P}(\bar{Y}_n \leq 0) \approx 50/100 = 50\%$. Or, if $\bar{Y}$ is in the interval $[-0.4, 0.4]$ in 70 of 100 samples, then $\mathrm{P}(-0.4 \leq \bar{Y} \leq 0.4) \approx 70\%$.

3.3.3 Bayesian and Frequentist Differences

The following makes explicit some of the differences between the Bayesian and frequentist approaches described above.

First, the frameworks treat different variables as random or non-random. The frequentist framework treats the population mean and other population features as non-random values, whereas it treats the data as random. For example, the population mean $\mu = \mathrm{E}(Y)$ is a non-random value, whereas an observation Y is a random variable. In contrast, the Bayesian framework treats (beliefs about) population features as random, whereas it treats the data as non-random values (the "after" view).

Second, due to this different treatment, the frameworks answer different types of questions, especially when quantifying uncertainty. The Bayesian framework answers questions about our beliefs after seeing the data. The frequentist framework answers questions about probabilities of seeing different features in the data, given the true population values.

Example 3.13 (Kaplan video)**.** Consider the question, "Given the observed data, what do I believe is the probability that the population mean is above 1/2?" This is a Bayesian question. Mathematically, if y is the "observed data," this question is commonly written as $\mathrm{P}(\mu > 1/2 \mid y)$, noting the conventional but confusing notation where μ represents beliefs. This question makes no sense from the frequentist perspective: either $\mu > 1/2$ or not; it cannot be "maybe," with some probability.

Example 3.14 (Kaplan video)**.** Consider the question, "Given the value of $\mu = \mathrm{E}(Y)$, what's the probability that the sample mean is above 1/2?" This is a frequentist question. Mathematically, this is usually written $\mathrm{P}(\bar{Y} > 1/2)$, or $\mathrm{P}_{\mu}(\bar{Y} > 1/2)$ to be explicit about the dependence on μ. The sample mean $\bar{Y}$ is a function of data, so it is treated as a random variable. This question makes no sense from the Bayesian perspective: we can see the data, so we can see either $\bar{Y} > 1/2$ or not; it cannot be "maybe," with some probability.

Interestingly, both frameworks can answer questions like $\mathrm{P}(\bar{Y} < \mu)$, but with different interpretations. The Bayesian answer interprets $\bar{Y}$ as a number (that we see in the data) and μ as a random variable representing our beliefs about the population mean. The frequentist answer interprets $\bar{Y}$ as the random variable (from the "before" view) and μ as the non-random population value.

Third, frequentist methods use only the data, whereas Bayesian methods can formally incorporate additional knowledge. In practice, though, even frequentist results should be interpreted in light of other knowledge. The difference is that this process is not formalized within the frequentist methodology itself. Unfortunately, many people do not combine frequentist results with other knowledge, instead interpreting frequentist results as if one single dataset contains the full, absolute truth of the universe; please do not do this!

In Sum: Bayesian & Frequentist

Frequentist: "before" view of data (random variables); assess methods' performance across repeated random samples from same population
Bayesian: "after" view of data (non-random); model beliefs (about population features) as random variables

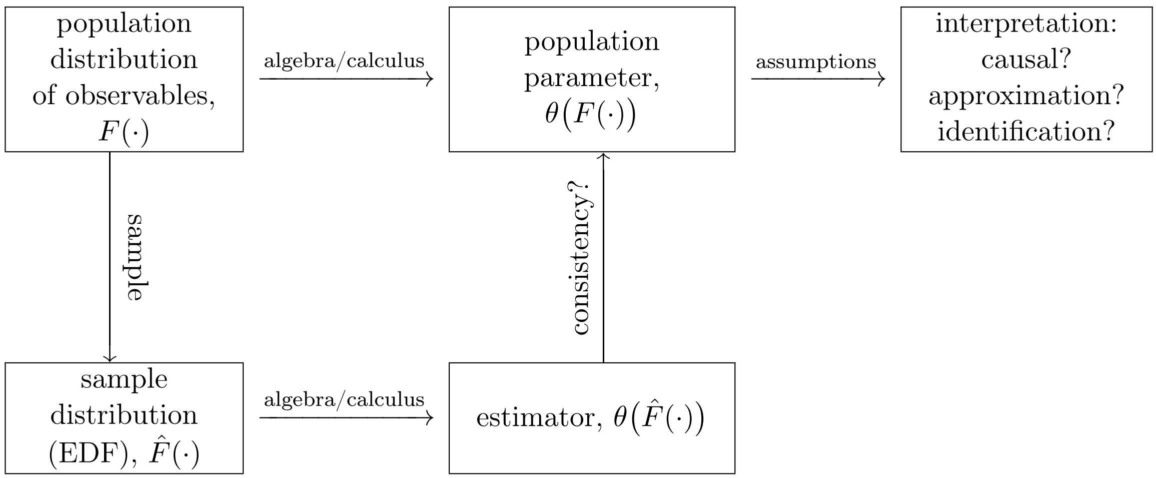


Figure 3.1: Map of (part of) the world of econometrics (Mercator projection).

3.4 Identification, Estimation, and Inference

Figure 3.1 shows one perspective of what (some) econometrics is about. Different parts of the "map" corresponds to identification and estimation.

Identification relates to the top-right of Figure 3.1. There are two different ways to think about identification. First, as in the map, imagine population parameter θ is a feature of the joint population distribution of observable variables, $F(\cdot)$. For example, maybe $\theta = \mathrm{E}(Y)$ is the mean of Y, or maybe $\theta = \mathrm{Cov}(Y, X)/\mathrm{Var}(X)$ is the slope of the linear projection of Y onto $(1, X)$. In some cases, this slope can be interpreted as the causal effect of X on Y. Identification can be understood in terms of the set of assumptions under which the population feature θ has this particular causal interpretation. Alternatively, imagine we define β as the causal effect of X on Y. This β is not a feature of the population distribution of (Y, X). However, β is **identified** if (under a set of **identifying assumptions**) it equals a feature of the population distribution of observables. More specifically, a parameter β is **point identified** $F(\cdot)$ uniquely determines the value of β. That is, if we somehow knew $F(\cdot)$, then we would also know the value of β. This book focuses on point identification of causal parameters.

Beyond our scope...

Parameters can also be **set identified** or **partially identified**, meaning that $F(\cdot)$ does not uniquely map to a single value of the parameter but rather a set of possible parameters. For example, maybe knowing $F(\cdot)$ lets us narrow down the possible values of β to the interval $[a, b]$, but $a < b$. For example, see Part VI of

Kaplan (2021).

Estimation relates to the bottom of Figure 3.1. Given identification, our object of interest is a feature of the population distribution of observables. In many cases, to get an estimator, we simply compute the same feature of the sample distribution, also called the empirical distribution, $\hat{F}(\cdot)$. This is called the **analogy principle** or **plug-in principle**. Other population parameters are defined as the solution to a population optimization problem, in which case the estimator solves the sample version of the problem. The OLS estimator can be thought of from both perspectives: it estimates the population parameter $\boldsymbol{\beta} = [\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\mathrm{E}(\boldsymbol{X}Y)$ by $\hat{\boldsymbol{\beta}} = [\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\boldsymbol{X}Y)$, or equivalently it estimates the population parameter $\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}} \mathrm{E}[(Y - \boldsymbol{X}'\boldsymbol{b})^2]$ by $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}} \widehat{\mathrm{E}}[(Y - \boldsymbol{X}'\boldsymbol{b})^2]$.

Inference is a vague word and also not well represented by Figure 3.1. People use inference in a variety of contexts with different meaning: Bayesian inference, causal inference, statistical inference, etc. In this book, it refers to methods (mostly confidence intervals) that quantify the statistical uncertainty about a certain population feature, which may not be the actual parameter of interest if the identifying assumptions are violated. For example, if you run `reg y x` in Stata, the confidence interval is for the population linear projection coefficients; even if you are interested in the causal effect of `x` on `y`, the confidence interval cannot account for your uncertainty about the identifying assumptions required for the linear projection slope to have a causal interpretation. However, sometimes there are ways to empirically assess certain identifying assumptions, as we will see.

3.5 General Equilibrium and Partial Equilibrium

This subsection is Section 4.3.3 of Kaplan (2022).

Another econometric dichotomy is between **general equilibrium** (GE) and **partial equilibrium** (PE) analysis. GE more ambitiously tries to model entire markets, sometimes multiple markets, whereas PE takes current market equilibria as given. The tradeoff is that the GE framework can analyze policies that change equilibria (i.e., that have **general equilibrium effects**), but it requires stronger assumptions to do so.

Example 3.15 (Kaplan video)**.** Imagine you were analyzing the impact of free public childcare on mothers' employment. A PE analysis would consider how mothers might respond to different childcare policies given the current prices of private childcare, current wages, etc. A GE analysis might further model the childcare and labor markets, to allow for the possible general equilibrium effects of public childcare policy on the prices in those markets. If there is a big expansion of free public childcare, then private childcares may indeed change their prices. If the expansion allows many mothers to enter the workforce, then the labor supply curve shifts out, which could lower wages. However, if the proposed changes to childcare policy are relatively small, then such GE effects may be negligible, and PE analysis may suffice.

The famous Lucas critique (Lucas, 1976) argues in part that macroeconomic policy analysis requires structural, GE models. Lucas writes (p. 41), "Given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models." That is, he says that if we want to guess how people and firms will behave in the future, under new macroeconomic policies, then we need to account for GE effects, which requires deeper, structural understanding and modeling of economic behavior.

In Sum: General & Partial Equilibrium Models

Partial equilibrium models treat prices and other market equilibria as fixed, whereas **general equilibrium models** allow markets to change.

3.6 Structural and Reduced-Form Approaches

This subsection is a shorter version of Section 4.3.2 of Kaplan (2022).

There are two general econometric approaches to learning about causality: the reduced-form approach, and the structural approach. Confusingly, the reduced-form approach is sometimes called **causal inference** even though the structural approach also aims to learn about causality. (Also confusingly, "reduced form" can refer to other related but different concepts.)

Both approaches consider **counterfactual** analysis, but in different ways. Broadly, a counterfactual is a universe that's different than our actual universe. Usually, the counterfactual universe is nearly identical to our actual universe except for one particular policy whose effect we want to learn.

The **reduced-form** approach tries to isolate causal effects by using comparisons that are either randomized or "as good as randomized." For example, **randomized** treatment means that individuals are randomly assigned to be treated or not, without regard to their characteristics. Hopefully, it is then appropriate to interpret the mean difference as the average effect of the treatment. "As good as randomized" means that although we did not explicitly randomly assign treatment, the actual assignment mechanism did not depend on individuals' characteristics anyway. More often this is (hoped to be) true after some other adjustment is made.

In contrast, the **structural approach** tries to more explicitly model the inner workings of causal systems. Structural models often come from economic theory, like decision-making or market equilibria models. The goal is to estimate such models' parameters, like elasticities, discount factors, risk aversion, and demand curves. There are different ways people define "structural," but I think a helpful definition is: a model that is invariant to a set of policies under consideration. If we are considering very large, macro-level policy changes, then we would need a relatively complex model, otherwise the policy changes could change the model (for example, through general equilibrium effects). If we are considering relatively small, micro-level policy changes, then a simpler model may suffice. Either way, the hope is that we can estimate the structural model and use it to guess the causal effect of each possible policy change.

The structural and reduced-form approaches have complementary advantages, and often both are helpful; for example, see the survey by Lewbel (2019). Structural models often require stronger (less realistic) assumptions, but in return they can analyze a wider variety of possible policies. Also, there can be relatively vague "structural" models (like in this book!), or relatively complex reduced-form models.

Example 3.16 (Kaplan video)**.** Imagine trying to learn how a retirement pension formula (i.e., how much money somebody gets paid after retiring, based on their years of experience, age, and salary history) affects the age at which a teacher decides to retire. A reduced-form analysis might compare the mean retirement age of teachers who joined a school in the year 1998 with the mean retirement age of teachers who joined in 1999, just after the formula was changed, hoping that the two groups of teachers are otherwise "as good as randomized." A structural analysis might explicitly model a teacher's retirement decision within an expected utility framework that "discounts" the value of future periods (like net present value). The structural analysis requires strong (maybe unrealistic) assumptions about things like the utility function and the distribution of unobserved variables. However, it can then evaluate the effect of hypothetical pension changes that may have never been implemented before, rather than only estimating the effect of the historical 1999 pension change.

Example 3.17. Imagine trying to learn about the effect of free public childcare on how much mothers work in the formal sector. A reduced-form analysis might estimate how much mothers work in cities that just opened such childcare centers last year compared to mothers in cities that plan to open them next year. The hope is that whether a city opens the childcare centers last year or next year is "as good as randomized," so that the mean difference in hours worked can be interpreted as the effect of the childcare (rather than the effect of something else that's different). A structural analysis might try to estimate an economic model of a mother's decision to work in the formal sector, including variables like the price of childcare, wages, and utility from different activities. Such a model requires strong assumptions (although "as good as randomized" may also be unrealistic!), but can then be used to evaluate the effects of a wide variety of hypothetical policies, not only the effect of the childcare centers that opened last year.

In Sum: Structural & Reduced-Form Approaches

Reduced-form: randomized or "as good as randomized" comparisons to isolate causality
Structural: more explicit economic models of causal relationships

3.7 Linear Regression

Discussion Question 3.6 (Model interpretation)**.** Interpret $Y = \beta_0 + \beta_1 X + U$. (As a concrete example: Y is wage, X is years of education.) In particular,

a) what does β_1 mean?
b) what does U mean?

The method of **ordinary least squares** (OLS) can be defined in multiple ways that each help illustrate a more general point.

The following notation is used in later chapters, too. Let Y be a scalar random variable that is the outcome of interest, like an individual's earnings or a state's traffic fatality rate. Let $\boldsymbol{X}$ be a column vector containing all the **regressors**, also known as **covariates** or **predictors** or **right-hand-side variables** or **independent variables**, usually with 1 as the first element, like $\boldsymbol{X} = (1, X_2, X_3, \ldots)'$. (Note that Wooldridge (2010) defines $\boldsymbol{X}$ as a row vector to avoid needing as many transpose symbols, but at the expense of needing to remember which vectors are columns vs. rows.) In this book, usually U is an unobserved scalar structural error term (with some causal meaning), whereas V is a statistical error term defined with respect to a linear projection or conditional mean.

3.7.1 Linear Projection

This subsection draws from Sections 7.3–7.5 of Kaplan (2022).

Fundamentally, OLS estimates the coefficients of the **linear projection** (LP) of Y onto $\boldsymbol{X}$. The LP is a population object:

$$\mathrm{LP}(Y \mid \boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\beta}, \tag{3.4}$$

where vector

$$\boldsymbol{\beta} = [\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\mathrm{E}(\boldsymbol{X}Y) \tag{3.5}$$

contains the **linear projection coefficients** (LPCs). This $\boldsymbol{\beta}$ is a feature of the population, i.e., it is a summary of the joint distribution of $(Y, \boldsymbol{X}')$. Assuming $\boldsymbol{\beta}$ is well-defined, iid sampling is sufficient for the OLS estimator $\hat{\boldsymbol{\beta}}$ to be consistent for $\boldsymbol{\beta}$, written $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$; details are below.

The **best linear predictor** (BLP) is another interpretation of the population LP. Often we think of "prediction" in terms of data, but here it is meant in the sense of trying to guess Y given $\boldsymbol{X}$ in the population. The "best" guess depends on the consequences of a wrong guess. For mathematical convenience, this is often quantified by squaring the difference between the guessed value and the true value Y, which is called **quadratic loss** (or L_2 loss). In BLP, "linear predictor" means a guess of the form $\boldsymbol{X}'\boldsymbol{g}$ (a linear combination of the predictors $\boldsymbol{X}$, where $\boldsymbol{g}$ is a non-random vector). It turns out that the LPC $\boldsymbol{\beta}$ solves

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{g}} \mathrm{E}[(Y - \boldsymbol{X}'\boldsymbol{g})^2]. \tag{3.6}$$

That is, among all possible predictions of the form $\boldsymbol{X}'\boldsymbol{g}$, the mean squared prediction error $\mathrm{E}[(\text{true} - \text{guess})^2]$ is minimized by setting $\boldsymbol{g} = \boldsymbol{\beta}$. Besides the caveat that quadratic loss may not reflect the actual consequences of our incorrect predictions, another caveat is that "best" does not mean "good": it could be that all predictions of the form $\boldsymbol{X}'\boldsymbol{g}$ are awful, and the BLP is merely the least bad.

Beyond our scope...

What if we replace quadratic loss with another loss function? If instead of squaring the error we take the absolute value, then we get the so-called "median regression" estimator. If more generally we use a tilted version of the absolute value function that allows positive errors to be worse (or better) than negative errors, then we get "quantile regression." For example, see Part II of Kaplan (2021).

The **best linear approximation** (BLA) is yet another interpretation of the LP. "Best" again refers to minimizing mean squared error (and again does not mean "good"!), and "linear" again refers to the functional form $\boldsymbol{X}'\boldsymbol{g}$ that takes a linear combination of $\boldsymbol{X}$. "Approximation" refers to approximation of the conditional mean $\mathrm{E}(Y \mid \boldsymbol{X})$. That is, the LPC $\boldsymbol{\beta}$ also solves

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{g}} \mathrm{E}\{[\mathrm{E}(Y \mid \boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{g}]^2\}. \tag{3.7}$$

This could also be written in terms of the **conditional mean function** (CMF), also called the **conditional expectation function** (CEF),

$$m(\boldsymbol{x}) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}). \tag{3.8}$$

Note that $m(\cdot)$ is a non-random function: it maps each possible non-random value $\boldsymbol{x}$ (lowercase) to the corresponding non-random scalar $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$, i.e., the mean Y among individuals in the subpopulation with $\boldsymbol{X} = \boldsymbol{x}$. Given this $m(\cdot)$, (3.7) can be written

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{g}} \mathrm{E}\{[m(\boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{g}]^2.\} \tag{3.9}$$

As noted above, OLS most fundamentally estimates the LP/BLP/BLA. The OLS estimator can be written as the sample analog of (3.5),

$$\hat{\boldsymbol{\beta}} = [\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\boldsymbol{X}Y). \tag{3.10}$$

Given iid sampling, sample moments converge in probability to the corresponding population moments by the weak law of large numbers (WLLN), so

$$\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}') \xrightarrow{p} \mathrm{E}(\boldsymbol{X}\boldsymbol{X}'), \quad \widehat{\mathrm{E}}(\boldsymbol{X}Y) \xrightarrow{p} \mathrm{E}(\boldsymbol{X}Y),$$

which can then be combined by the continuous mapping theorem (again assuming everything is well-defined):

$$\hat{\boldsymbol{\beta}} = [\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\boldsymbol{X}Y) \xrightarrow{p} [\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\mathrm{E}(\boldsymbol{X}Y) = \boldsymbol{\beta}.$$

The above **sample analog** form of the OLS estimator can be derived from the "least squares" definition that mirrors (3.6),

$$\hat{\boldsymbol{\beta}} \equiv \arg\min_{\boldsymbol{b}} \widehat{\mathrm{E}}[(Y - \boldsymbol{X}'\boldsymbol{b})^2] = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i'\boldsymbol{b})^2. \tag{3.11}$$

The objective function is clearly convex (satisfying the second-order condition), so the (unique global) minimizer solves the first-order condition

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{b}}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i'\boldsymbol{b})^2\bigg|_{\boldsymbol{b}=\hat{\boldsymbol{\beta}}} = \frac{1}{n}\sum_{i=1}^{n}2\boldsymbol{X}_i(Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}) = \frac{2}{n}\sum_{i=1}^{n}(\boldsymbol{X}_iY_i - \boldsymbol{X}_i\boldsymbol{X}_i'\hat{\boldsymbol{\beta}}).$$

Dividing by 2 and solving for $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_iY_i = [\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\boldsymbol{X}Y).$$

Just as we can interpret the population $\boldsymbol{\beta}$ in terms of LP, BLP, or BLA, we can also interpret the estimator in terms of the sample minimization problem (3.11) that parallels the BLP population minimization problem, or as the sample analog (3.10) of the population LPC expression in (3.5).

Often the linear projection model is written in **error form**, but note that the above does not fundamentally require any such "error term." Nonetheless, sometimes it is convenient to define the linear projection error V as the difference between the true Y and the linear projection:

$$V \equiv Y - \mathrm{LP}(Y \mid \boldsymbol{X}) = Y - \boldsymbol{X}'\boldsymbol{\beta}. \tag{3.12}$$

Given this definition of V, it follows automatically that

$$\mathrm{E}(\boldsymbol{X}V) = \mathbf{0},$$

meaning every element of the vector $\mathrm{E}(\boldsymbol{X}V)$ is zero. This is not an assumption: it is a property that follows from the definition of V as the linear projection error.

3.7.2 Conditional Mean Function

Linear CMF is LP/BLP/BLA

While OLS most fundamentally estimates the LP/BLP/BLA, under stronger assumptions it can estimate the CMF defined in (3.8). From (3.9), if the true CMF happens to have the linear functional form $m(\boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\gamma}$ for some non-random vector $\boldsymbol{\gamma}$, then

$$m(\boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{\gamma} - \boldsymbol{X}'\boldsymbol{b}$$

can be set to zero by setting $\boldsymbol{b} = \boldsymbol{\gamma}$, in which case the entire RHS is zero and thus the global minimum because the RHS is non-negative (due to the squaring). Thus, the RHS is $\boldsymbol{\gamma}$ (defined above to be the CMF coefficient vector), and the LHS is the LPC $\boldsymbol{\beta}$, so $\boldsymbol{\beta} = \boldsymbol{\gamma}$ and the LP and CMF are equal. That is, if the CMF is "linear" in the sense of having the same functional form as the LP, then the LP and CMF are equal, so we can interpret the OLS estimand (which fundamentally is the LP) as the CMF. In some cases, this is always true, like if $\boldsymbol{X} = (1, X)'$ where X is a dummy variable (only takes value 0 or 1).

Beyond our scope...
There are ways to estimate the CMF without requiring that you guess the exact functional form ahead of time, called **nonparametric regression**; for example, see Part V of Kaplan (2021).

CMF in Error Form

Like the LP, the CMF can also be written in error form. Parallel to the LP error defined in (3.12), the CMF error is defined as

$$V \equiv Y - m(\boldsymbol{X}), \tag{3.13}$$

where $m(\cdot)$ is the CMF defined in (3.8). The property $\mathrm{E}(V \mid \boldsymbol{X}) = 0$ follows from the definition; it is not an additional assumption. That is, if we write

$$Y = m(\boldsymbol{X}) + V \tag{3.14}$$

with $m(\cdot)$ the CMF from (3.8), then it automatically follows that

$$\mathrm{E}(V \mid \boldsymbol{X}) = \mathrm{E}(Y - m(\boldsymbol{X}) \mid \boldsymbol{X}) = \mathrm{E}(Y \mid \boldsymbol{X}) - m(\boldsymbol{X}) = 0, \tag{3.15}$$

where the first equality plugs in the definition in (3.13), the second uses the linearity property of $\mathrm{E}(\cdot)$, and the third uses the definition of $m(\cdot)$.

CMF vs. Conditional Mean

One common confusion is the difference between $m(\boldsymbol{x})$ and $m(\boldsymbol{X})$. The former is a non-random function evaluated at a non-random value, hence $m(\boldsymbol{x})$ is a non-random value, like 7 or -1.1. The latter is a non-random function evaluated at a random value $\boldsymbol{X}$, hence $m(\boldsymbol{X})$ is also a random variable. For example, imagine scalar X with $\mathrm{P}(X = 0) = \mathrm{P}(X = 1) = 0.5$, and $m(x) = x + 2$; then $\mathrm{P}(m(X) = 2) = \mathrm{P}(X = 0) = 0.5$ and $\mathrm{P}(m(X) = 3) = \mathrm{P}(X = 1) = 0.5$, showing that $m(X)$ is a random variable, whereas $m(0) = 2$ and $m(1) = 3$ are non-random values.

Similarly, $\mathrm{E}(V \mid \boldsymbol{X})$ is a random variable, whereas $\mathrm{E}(V \mid \boldsymbol{X} = \boldsymbol{x})$ is a non-random value. The expression $\mathrm{E}(V \mid \boldsymbol{X}) = 0$ means that zero is the only possible value the random variable takes. Equivalently, we could also write $\mathrm{E}(V \mid \boldsymbol{X} = \boldsymbol{x}) = 0$ for all possible values $\boldsymbol{x}$, which may be easier to understand.

CMF with Binary Regressor

Consider the special case with binary X. Let $m(x) = \beta_0 + \beta_1 x$. Then we can solve for the parameters from $m(0) = \beta_0$ and $m(1) = \beta_0 + \beta_1$, implying $\beta_1 = m(1) - m(0)$. That is, the intercept β_0 is the conditional mean of Y when $X = 0$, and the slope β_1 is the difference in conditional means of Y between $X = 1$ and $X = 0$.

3.7.3 Causal Interpretation

Under additional assumptions, the LP or CMF can have a causal interpretation. This is left to later chapters.

3.8 Economic Significance

3.8.1 Basic Idea

The term **economic significance** refers to the magnitude of an estimated parameter. An estimate is not economically significant if it is "economically" negligible (not meaningfully different than zero). "Economically" just means "for real-world purposes," like whether it is important to consider for policy purposes. One way to think about this is: would you personally care about the difference? For example, imagine $\hat{\theta}$ estimates the effect on your final exam score of studying an additional hour per week. Would you care about having a final exam score that's $\hat{\theta}$ percentage points higher? If $\hat{\theta} = 0.01$, then no; if $\hat{\theta} = 50$, then yes. Of course, it's a continuum, so somewhere between "yes" and "no" are varying degrees of "maybe," corresponding to varying degrees of moderate economic significance.

Example 3.18. Would you care if you had $\hat{\theta} = 2$ additional years of education? This is a lot, like an entire master's degree, so presumably you would indeed care.

3.8.2 Units of Measure

It is very important to consider units of measure. For example, imagine the estimated effect on income is $\hat{\theta} = 10$; is that economically significant? If the units are dollars per hour, then yes; if it's dollars per year, then no; if it's thousands of dollars per month, then yes; etc.

It is also very important to consider realistic policy changes (which usually requires paying attention to the units of X). For example, imagine your estimated $\hat{\theta}$ is the effect of a one-unit increase in the proportion of the state budget allocated to higher education. If the current proportion is 0.08 (meaning 8%), then a realistic policy change would be something like 0.02 units. A one-unit increase would mean changing from 0% to 100% of the budget spent on higher education. Even if $\hat{\theta}$ looks economically significant, maybe $0.02\hat{\theta}$ does not.

3.8.3 Log Models

In addition to units of measure, coefficient interpretations depend on whether a variable enters the model in levels or in logs. In economics, "log" always refers to the natural log.

The interpretations of different log models are detailed in Section 8.1 of Kaplan (2022). XXX SUMMARIZE

3.9 Quantifying Uncertainty

See also Section 3.7 of Kaplan (2022), as well as Section 3.8 ("Quantifying Uncertainty: Misinterpretation and Misuse").

While an estimator provides a best guess about the true population value given the data (roughly speaking), we usually also want a sense of our uncertainty about the true value. The most common frequentist methods to quantify uncertainty are confidence intervals and p-values from hypothesis tests. This book focuses on confidence intervals because most econometricians agree they are less likely to be misinterpreted or misused (than p-values). Additionally, in many settings, with large n the frequentist confidence interval is very similar to the Bayesian credible interval. Hopefully you have already learned the basics of hypothesis testing and p-values (because they are still reported, and occasionally still useful), like that an estimate $\hat{\theta}$ is **statistically significant** at level α if the p-value for $H_0\colon \theta = 0$ is below α, or that failing to reject such a null hypothesis should not be interpreted as our best guess being $\theta = 0$. For example, if $\hat{\theta} = 0.1$ and $p = 0.33$, then we would not be surprised to see such a dataset if indeed $\theta = 0$, but we would also not be surprised to see such a dataset if $\theta = 0.2$, etc.

A **confidence interval** (CI) only quantifies uncertainty due to random sampling, not uncertainty about identifying assumptions. For example, a CI for the linear projection slope accounts only for the uncertainty due to having finite sample size n, not due to uncertainty about the true CMF being linear, nor due to uncertainty about the CMF slope having a causal interpretation. This can be misleading. If we have very large n, then our CI will be very short (because we have very little uncertainty about the LPC), even if we are very uncertain about the CMF being linear or having a causal interpretation. For this reason, it is useful to know the fundamental population parameter that a particular estimator is consistent for, like how OLS is fundamentally consistent for the LPC.

A CI provides a range of values that should contain the true value with high probability. Recall that from the frequentist perspective, "probability" is from the before-sampling perspective, and the CI is random (because it depends on the observations, which are modeled as random), whereas the true population value is non-random. That is, a CI can be seen as a procedure such that (before sampling) we have a high probability of randomly sampling a dataset for which the CI contains the true value. This probability is called the **coverage probability** (CP). That is, given population value θ and CI $[\hat{L}, \hat{U}]$, where the lower and upper endpoints $\hat{L}$ and $\hat{U}$ are computed from data (thus random variables), the coverage probability is

$$\mathrm{CP} \equiv \mathrm{P}(\hat{L} \leq \theta \leq \hat{U}) = \mathrm{P}(\theta \in [\hat{L}, \hat{U}]). \tag{3.16}$$

The **confidence level** or **nominal coverage probability** is the desired coverage probability. Usually a CI is justified by an asymptotic argument such that asymptotically, its coverage probability equals the nominal level. (Sometimes the asymptotic coverage probability is only shown to be greater than or equal to the nominal level.) However, for finite n, the coverage probability may be higher or lower than desired. If it is higher than desired, then the CI is "too conservative": hypothetically, it could be shortened and still achieve the desired coverage probability. Lower than desired CP is usually considered even worse: we are over-confident about how precise our estimates are. Of course, even with a 95% CI, our CI fails to include the true value 5% of the time, so we should never be too confident anyway. This is why replication is an important part of any science (although that begs the question of whether economics is truly a science!).

Coverage probability can also be interpreted in terms of repeated sampling. For example, if we have a 90% probability of randomly sampling a dataset for which the CI contains the true value, and we randomly sample 100 datasets, then in roughly 90 of the 100 datasets the CI should contain the true value.

Instead of a binary label of "statistically significant at a 5% level" whenever $p < 0.05$, it is more helpful to look at the full range of possible population values included in the CI when quantifying uncertainty. At minimum: consider the economic significance of the lowest value in the CI, the estimated value, and the highest value in the CI. If a confidence level $100(1-\alpha)\%$ CI does not contain zero, then it is "statistically significant at level α," but that is usually not the most helpful statement to make. For example, if a 95% CI is $[0.1, 1.1]$, then there is statistical significance at a 5% level.

Discussion Question 3.7 (salary increase significance)**.** Imagine you compute a 95% CI of $[4.1, 5.9]$ around your estimated annual salary effect of $\hat{\theta} = 5$ dollars per year. Are these results statistically significant? Are they economically significant? Explain. Hint: would you care if your annual salary increased by $\hat{\theta} = 5$ dollars per year?

Discussion Question 3.8 (significance: distance and education)**.** Let Y be years of education, and let X be distance from someone's childhood home to the nearest college or university, measured in kilometers ($1\,\text{km} = 0.6\,\text{mi}$). Let θ be the causal effect of X on Y. You think you found an "as good as randomized" natural experiment, from which you estimate $\hat{\theta} = -0.03$. You compute a 95% CI of $[-0.05, -0.01]$.

a) How economically significant is the point estimate of -0.03? Hint: consider the units.
b) Is this statistically significant at a 5% level?
c) More generally, discuss your CI and uncertainty.

There are many possible ways to misinterpret or misuse confidence intervals (or p-values), including the following (not exhaustive!).

- Multiple testing: if you take enough random samples, or test enough different hypotheses in the same sample, you will eventually get a "statistically signficant" result; for example, see this insightful comic (`xkcd.com/882`) that illustrates the **multiple testing problem** (or **multiple comparisons problem**), or this video.

- Publication bias: if statistically significant results are more likely to be published, then it's similar to the multiple testing problem in the linked comic, where we only read about the one significant result but not the 19 not-significant results.

- Assumptions: a CI may not be valid if it is based on iid sampling but the actual data were not sampled iid; and the CI does not account for additional interpretations of the population value based on identifying assumptions.

- Frequentist results may be misinterpreted as Bayesian, like a p-value being misinterpreted as the probability that the null hypothesis is true.

- Unlikely events happen: even if you only run one test on one dataset with confidence level 99%, your dataset may be in the unlucky 1% for which the true value is outside the CI.

Example 3.19 (Kaplan video)**.** Your friend claims to have magical powers. You have a deck of playing cards; you repeatedly draw a card (without showing it) and ask your friend to guess whether the card is black or red. You record the data and compute a 90% CI for your friend's probability p of guessing correctly. Random guessing would yield $p = 0.5$, but your CI is $[0.52, 0.61]$, all values above 0.5. Your friend's interpretation is that statistics have now proved true the claim of magical powers. However, you think it was just luck and ask to gather more data. Indeed, the new dataset's 90% CI is $[0.44, 0.51]$. You try another few datasets, and those CIs also contain 0.5. It seems the first result was simply luck, not magic.

Discussion Question 3.9 (frequentist or Bayesian?)**.** For each of the following, say whether it is a frequentist question, Bayesian question, neither, or both; if both, explain the two possible interpretations. Hint: use Section 3.3 as well as Section 3.9.

a) What's the probability that the current natural unemployment rate in the U.S. is between 4.5% and 7.5%?
b) Can we create a diagnostic tool for our company's daily website traffic data to identify whether it's normal or has been hacked, limiting the rate of falsely reporting "hacked" on normal days to only 1% of normal days?
c) What is the probability that the true unemployment rate is within 1 percentage point of the estimated unemployment rate?
d) Is the positive estimate $\hat{\theta} > 0$ primarily due to the income effect or substitution effect?

3.10 Quantifying Accuracy of an Estimator

This section is mostly from Section 3.6 of Kaplan (2022).

From the frequentist perspective, an estimator's accuracy can be quantified by comparing features of its sampling distribution to the true population value. The **sampling distribution** views the estimator from the before-sampling perspective; for intuition, you can imagine taking 1000 random samples and plotting a histogram of the estimated values. Bias is an important, commonly mentioned property, but it is not sufficient to quantify accuracy. Mean squared error better quantifies accuracy.

Throughout, let θ be the population parameter estimated by $\hat{\theta}_n$; for example, $\theta = \mathrm{E}(Y)$ and $\hat{\theta}_n = \bar{Y}_n$.

3.10.1 Bias

Definitions

The **bias** of $\hat{\theta}_n$ compares the mean of its sampling distribution to the true population θ. Mathematically,

$$\mathrm{Bias}(\hat{\theta}_n) \equiv \mathrm{E}(\hat{\theta}_n) - \theta. \tag{3.17}$$

The bias captures if the estimator systematically differs from θ in a particular direction, i.e., how wrong the average $\hat{\theta}_n$ is.

There are four types of bias:

$$\begin{aligned}
\textbf{upward bias (positive bias)}\text{:}\ & \mathrm{E}(\hat{\theta}_n) > \theta, \\
\textbf{downward bias (negative bias)}\text{:}\ & \mathrm{E}(\hat{\theta}_n) < \theta, \\
\textbf{attenuation bias (bias toward zero)}\text{:}\ & 0 < \frac{\mathrm{E}(\hat{\theta}_n)}{\theta} < 1, \text{so } |\mathrm{E}(\hat{\theta}_n)| < |\theta|, \\
\text{bias away from zero:}\ & \frac{\mathrm{E}(\hat{\theta}_n)}{\theta} > 1, \text{so } |\mathrm{E}(\hat{\theta}_n)| > |\theta|.
\end{aligned}$$

An estimator is **unbiased** if its bias is zero. Using (3.17),

$$\mathrm{Bias}(\hat{\theta}) = 0 \iff \mathrm{E}(\hat{\theta}) = \theta, \tag{3.18}$$

where symbol $\iff$ can be read as "is equivalent to" (see Chapter 2).

Example 3.20 (Kaplan video)**.** With iid sampling, the sample mean is an unbiased estimator of the population mean. The estimator is $\hat{\theta}_n = \bar{Y}_n$, and the population parameter is $\theta = \mathrm{E}(Y)$. With $n = 1$, $\bar{Y}_1 = Y_1$, so $\mathrm{E}(\bar{Y}_1) = \mathrm{E}(Y_1) = \mathrm{E}(Y)$. With $n = 2$,

$$\mathrm{E}[\bar{Y}_2] = \mathrm{E}[(1/2)Y_1 + (1/2)Y_2] = \overbrace{(1/2)\,\mathrm{E}(Y_1)}^{\mathrm{E}(Y)/2} + \overbrace{(1/2)\,\mathrm{E}(Y_2)}^{\mathrm{E}(Y)/2} = \mathrm{E}(Y), \tag{3.19}$$

using the linearity property of $\mathrm{E}(\cdot)$. Similar derivations hold for any n, so $\mathrm{E}(\bar{Y}_n) = \mathrm{E}(Y)$, thus the bias is zero given (3.18).

Example 3.21 (Kaplan video)**.** The estimator $\hat{\theta}_n = \bar{Y}_n + 1$ has positive bias for the mean $\mathrm{E}(Y)$: $\mathrm{E}(\hat{\theta}_n) = \mathrm{E}(\bar{Y}_n + 1) = \mathrm{E}(\bar{Y}_n) + 1 = \mathrm{E}(Y) + 1 > \mathrm{E}(Y)$. The estimator $\hat{\theta}_n = \bar{Y}_n - 2$ has negative bias for the mean $\mathrm{E}(Y)$: $\mathrm{E}(\hat{\theta}_n) = \mathrm{E}(\bar{Y}_n - 2) = \mathrm{E}(\bar{Y}_n) - 2 = \mathrm{E}(Y) - 2 < \mathrm{E}(Y)$. The estimator $\hat{\theta}_n = 0.5\bar{Y}_n$ has attenuation bias for the mean $\mathrm{E}(Y)$: $\mathrm{E}(\hat{\theta}_n) = \mathrm{E}(0.5\bar{Y}_n) = 0.5\,\mathrm{E}(\bar{Y}_n) = 0.5\,\mathrm{E}(Y)$, so $0 < [\mathrm{E}(\hat{\theta}_n)/\,\mathrm{E}(Y)] = 0.5 < 1$.

Insufficiency of Bias to Quantify Accuracy

Bias alone does not fully quantify accuracy. That is, if you only consider bias when choosing between two possible estimators, then you may be fooled into choosing the worse estimator.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two different estimators of the same unknown parameter θ. Here, the subscripts 1 and 2 do not indicate n but just that the estimators are different. For simplicity, let $\theta = 0$. The first estimator's distribution is

$$\mathrm{P}(\hat{\theta}_1 = -100) = \mathrm{P}(\hat{\theta}_1 = 100) = 1/2. \tag{3.20}$$

The second estimator's distribution is

$$\mathrm{P}(\hat{\theta}_2 = 1) = 1. \tag{3.21}$$

The first estimator has smaller bias. The estimators' means are

$$\mathrm{E}(\hat{\theta}_1) = (1/2)(-100) + (1/2)(100) = 0, \quad \mathrm{E}(\hat{\theta}_2) = (1)(1) = 1. \tag{3.22}$$

Thus, recalling $\theta = 0$, the bias of each estimator is

$$\mathrm{Bias}(\hat{\theta}_1) = \mathrm{E}(\hat{\theta}_1) - \theta = 0 - 0 = 0, \quad \mathrm{Bias}(\hat{\theta}_2) = \mathrm{E}(\hat{\theta}_2) - \theta = 1 - 0 = 1. \tag{3.23}$$

Estimator $\hat{\theta}_1$ is unbiased, whereas $\hat{\theta}_2$ has upward bias.

But intuitively, $\hat{\theta}_2$ is much better. It always differs from the true θ by only 1, whereas $\hat{\theta}_1$ always differs by 100, which is much worse. That is, regardless of the dataset, $\hat{\theta}_2$ is always 100 times closer than θ_1 to the true $\theta = 0$. This illustrates how bias alone does not properly quantify our preferences: it tells us to prefer $\hat{\theta}_1$ (lower bias) when in fact we strongly prefer $\hat{\theta}_2$ (always much closer to θ).

3.10.2 Mean Squared Error

$\Longrightarrow$ Kaplan video: MSE Examples

The **mean squared error** (MSE) is a more complete measure of "how bad" an estimator is. The idea is analogous to using quadratic loss for prediction as in (3.6). Among other possible loss functions, this is most common and generally reasonable. MSE is mean quadratic loss:

$$\mathrm{MSE}(\hat{\theta}) \equiv \mathrm{E}[L_2(\hat{\theta},\theta)] = \mathrm{E}[(\hat{\theta}-\theta)^2]. \tag{3.24}$$

Continuing the example, our intuitive preference for $\hat{\theta}_2$ over $\hat{\theta}_1$ is supported by MSE. Because MSE measures "how bad" an estimator is, $\hat{\theta}_2$ being "better" means it has lower MSE. Specifically,

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_1) &= \mathrm{E}[(\hat{\theta}_1-\theta)^2] = (1/2)(-100-0)^2 + (1/2)(100-0)^2 = 10{,}000,\\ \mathrm{MSE}(\hat{\theta}_2) &= \mathrm{E}[(\hat{\theta}_2-\theta)^2] = (1)(1-0)^2 = 1.\end{aligned}$$

This matches our intuition: $\hat{\theta}_2$ is much better than $\hat{\theta}_1$ because it has much lower MSE.

MSE can also be decomposed into variance plus squared bias. The variance is

$$\mathrm{Var}(\hat{\theta}) \equiv \mathrm{E}[(\hat{\theta}-\mathrm{E}(\hat{\theta}))^2]. \tag{3.25}$$

(The square root of this is the standard deviation, also called the "standard error" of the estimator $\hat{\theta}$.) Skipping the math, using the bias and variance definitions in (3.17) and (3.25),

$$\mathrm{E}[(\hat{\theta}-\theta)^2] = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta})]^2. \tag{3.26}$$

All else equal, larger bias is bad, but it's also bad to have very high and very low estimates across datasets (large variance and "standard error") even if they happen to average to θ.

Example 3.22 (Kaplan video)**.** Continue the previous example, but instead of assuming $\theta = 0$, let

$$\mathrm{P}(\hat{\theta}_1 = \theta - 100) = \mathrm{P}(\hat{\theta}_1 = \theta + 100) = 1/2, \quad \mathrm{P}(\hat{\theta}_2 = \theta + 1) = 1. \tag{3.27}$$

The MSEs are the same as before because the θ cancels out:

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_1) &= \mathrm{E}[(\hat{\theta}_1-\theta)^2] = (1/2)(\theta-100-\theta)^2 + (1/2)(\theta+100-\theta)^2 = 10{,}000,\\ \mathrm{MSE}(\hat{\theta}_2) &= \mathrm{E}[(\hat{\theta}_2-\theta)^2] = (1)(\theta+1-\theta)^2 = 1.\end{aligned} \tag{3.28}$$

Example 3.23 (Kaplan video)**.** Imagine we know the bias and variance of two estimators, but not the full sampling distributions. This is still sufficient to compute MSE using (3.26). For example, let

$$\mathrm{Bias}(\hat{\beta}_1) = 1, \mathrm{Var}(\hat{\beta}_1) = 16, \quad \mathrm{Bias}(\hat{\beta}_2) = 10, \mathrm{Var}(\hat{\beta}_2) = 9. \tag{3.29}$$

Plugging these into (3.26),

$$\mathrm{MSE}(\hat{\beta}_1) = 1^2 + 16 = 17, \quad \mathrm{MSE}(\hat{\beta}_2) = 10^2 + 9 = 109. \tag{3.30}$$

According to MSE, $\hat{\beta}_1$ is better because it has lower MSE ("less bad") than $\hat{\beta}_2$. In this case, although $\hat{\beta}_1$ has larger variance, its bias is enough smaller than its overall MSE is also smaller.

Discussion Question 3.10 (estimator MSE)**.** Consider three estimators of the population mean $\mu = \mathrm{E}(Y)$, and their three sampling distributions: $\hat{\mu}_1 \sim \mathrm{N}(\mu, 25)$, $\hat{\mu}_2 \sim \mathrm{N}(\mu+3, 16)$, and $\hat{\mu}_3 \sim \mathrm{N}(\mu+2, 9)$, i.e., the sampling distributions of the three estimators are all normal distributions with respective means μ, $\mu+3$, and $\mu+2$, and respective variances 25, 16, and 9.

a) Compute the MSE of each estimator.
b) Rank the three estimators from best to worst, in terms of MSE.

3.10.3 Consistency and Asymptotic MSE

At the intuitive level, an estimator is **consistent** if in "large" samples (large n), there is a "high" probability of the estimator being "close" to the true value. This is similar to the idea of "probably approximately correct" in computer science: estimator $\hat{\theta}_n$ is "consistent" if with large n it is "probably approximately correct." Unfortunately, there are usually no precise quantitative definitions of "large," "high," and "close."

If $\hat{\theta}_n$ is not consistent, then it has **asymptotic bias**: even with infinite data, the estimator would still be biased. One way to formally define asymptotic bias is

$$\mathrm{AsyBias}(\hat{\theta}_n) \equiv \operatorname*{plim}_{n\to\infty} \hat{\theta}_n - \theta. \tag{3.31}$$

Analogous to "unbiasedness" being "zero bias," here "consistency" is "zero asymptotic bias": roughly speaking, with a large dataset, there is very little bias. There are the same four types of asymptotic bias as bias: upward/positive, downward/negative, attenuation, and away from zero.

It is also possible to compare approximate (asymptotic) mean squared error by comparing asymptotic distributions. Again, lower is better, and it depends on both bias and variance components. For two consistent estimators, this reduces to comparing asymptotic variance. For example, if $\sqrt{n}(\hat{\theta}_1 - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma_1^2)$ and $\sqrt{n}(\hat{\theta}_2 - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma_2^2)$, then we prefer estimator $\hat{\theta}_1$ (and call it more **efficient** than $\hat{\theta}_2$) iff $\sigma_1 < \sigma_2$.

Beyond our scope...

In contexts like nonparametric regression, there is also an important bias term, even asymptotically, and procedures are designed to try to minimize the asymptotic MSE; for example, see Chapter 18 ("Model Selection") of Kaplan (2021).

Chapter 4

Identification by Independence

Unit learning objectives for this chapter

4.1. Explain mathematically and verbally how an independence condition can achieve identification, in both structural and potential outcomes models. [TLOs 2 and 3]

4.2. In real-world examples, provide reasons why the key identifying assumption probably does (not) hold. [TLO 4]

To develop intuition and vocabulary, this chapter explains identification in the simplest structural and potential outcomes models.

Some material is from Chapters 4 and 6 of Kaplan (2022). Sections 21.1–3 of Wooldridge (2010) cover some of the same topics as Section 4.1.

4.1 Average Treatment Effect

First, the potential outcomes framework and notation are introduced. Then, the average treatment effect is defined, after which identification results are given.

4.1.1 Potential Outcomes

$\Longrightarrow$ Kaplan video: Potential Outcomes and the ATE

This subsection is a shorter version of Section 4.4.1 of Kaplan (2022).

The **potential outcomes framework** is also called the **Neyman–Rubin causal model** after its two earliest contributors (although sometimes Neyman's name is dropped). It is popular not only in economics, but statistics, medicine, political science, and other fields.

The terms **treatment** and **treatment effect** just refer to any variable and its causal effect on another variable. In English, usually "treatment" makes us think narrowly about medicine (or lumber. . . and facials?), but it can be anything. For example, the "treatment" could be a job training program, and the "treatment effect" is the causal effect of the program on a person's wage. Or, a treatment could be going to a charter school (instead of public school). Another treatment could be a policy or law, like a higher sales tax, or a certain labor law.

As throughout this book, "individual" can mean a firm, county, school, etc.

Imagine two parallel universes. The universes are identical except for one difference: whether or not an individual is treated. The individual's outcome in the universe without treatment is their **untreated potential outcome**, and the individual's outcome in the universe with treatment is their **treated potential outcome**.

Notationally, Y^t represents the treated potential outcome and Y^u the untreated potential outcome. Elsewhere, often Y_1 and Y_0 represent the treated and untreated potential outcomes, or $Y(1)$ and $Y(0)$.

Potential outcomes Y^u and Y^t are not always observable. Often, if an individual is untreated in our universe, then we can observe her untreated potential outcome Y^u, but not her Y^t; conversely, if she is treated, then we observe Y^t but not Y^u. This partial observability makes causal inference more difficult than description or prediction.

Example 4.1 (Kaplan video)**.** Imagine one universe where a student wins the lottery to enter a popular charter school, and another universe where the student remains in the conventional public school. Potential outcomes Y^t and Y^u are dummy (binary) variables for whether or not the student eventually graduated from college in each respective universe. Again, in our universe, we can observe Y^t if the student wins the lottery and Y^u if not, but we cannot observe both.

4.1.2 Treatment Effects

This subsection is a shorter version of Section 4.4.2 of Kaplan (2022).

The difference $Y^t - Y^u$ between an individual's two potential outcomes is that individual's **treatment effect**. Just as different individuals can have different (Y^u, Y^t), individuals can have different treatment effects $Y^t - Y^u$; i.e., individuals can be affected differently by the same treatment. The fancy term for people being different is **heterogeneity**, more specifically here "treatment effect heterogeneity."

Example 4.2 (Kaplan video)**.** In the charter school example (Example 4.1), $Y^t - Y^u$ is the treatment effect of the charter school on college graduation. That is, it is the difference between the college graduation outcomes in the charter school universe and the public school universe. Because the outcome is binary (1 if graduate college, 0 if don't), there are only four possible values of (Y^u, Y^t) (student types) and only three possible treatment effect values: $Y^t - Y^u = 1$ if the student graduates in the charter school universe ($Y^t = 1$) but not the public school universe ($Y^u = 0$); $Y^t - Y^u = -1$ if they only graduate in the public school universe ($Y^u = 1$) but not the charter school universe ($Y^t = 0$); and $Y^t - Y^u = 0$ if they graduate either in both universes ($Y^t = Y^u = 1$) or neither ($Y^t = Y^u = 0$). This is seen in the later example of Table 4.1.

In economics, where many systems are interrelated, sometimes it's difficult merely to specify which "effect" we care about. For example, consider racial differences in salary. In the parallel universe that's "identical" except for the individual's race, does "identical" include having the same job at the same firm? Or does it allow for an effect of race on hiring? Does it allow for an effect on educational opportunities, or an effect on family background (parents' education, wealth, etc.)? There is no "right" or "wrong" specification, but each answers a different question.

In Sum: Causality in Potential Outcomes Framework

Treatment effect: the difference in outcomes between parallel universes identical except for treatment

4.1.3 Average Treatment Effect

⟹ Kaplan video: Potential Outcomes and the ATE (again)

This subsection is a shorter version of Section 4.5 of Kaplan (2022).

Although the full distribution of potential outcomes (Y^u, Y^t) contains the most information, usually only certain summary features are studied; here, we focus on the mean.

The **average treatment effect** (ATE) is $\mathrm{E}(Y^t - Y^u)$. "Average" refers to the population mean, while "treatment effect" refers to $Y^t - Y^u$. Thus, the ATE may be interpreted as the probability-weighted average (mean) of all possible individual treatment effects in the population. Another name for the ATE is the **average causal effect** (ACE), but I use ATE to emphasize that this concept is from the potential outcomes framework.

The ATE has another interpretation. Using the linearity of the expectation operator,

$$\mathrm{ATE} \equiv \mathrm{E}(Y^t - Y^u) = \mathrm{E}(Y^t) - \mathrm{E}(Y^u). \tag{4.1}$$

Here, $\mathrm{E}(Y^t)$ is the mean treated potential outcome, and $\mathrm{E}(Y^u)$ is the mean untreated potential outcome. This could be interpreted as "the treatment effect on the mean outcome": treatment causes the mean outcome to change from $\mathrm{E}(Y^u)$ to $\mathrm{E}(Y^t)$.

Example 4.3 (Kaplan video)**.** Table 4.1 shows a numerical version of the charter school example. The four student "types" refer to the four possible values of (Y^u, Y^t), and each type has its own probability. Given the probabilities, the mean untreated outcome $\mathrm{E}(Y^u)$, mean treated outcome $\mathrm{E}(Y^t)$, and ATE $\mathrm{E}(Y^t - Y^u)$ are

$$\mathrm{E}(Y^u) = (0.3)(0) + (0.3)(0) + (0.1)(1) + (0.3)(1) = 0.4, \tag{4.2}$$

$$\mathrm{E}(Y^t) = (0.3)(0) + (0.3)(1) + (0.1)(0) + (0.3)(1) = 0.6, \tag{4.3}$$

$$\mathrm{E}(Y^t - Y^u) = (0.3)(0) + (0.3)(1) + (0.1)(-1) + (0.3)(0) = 0.2. \tag{4.4}$$

To verify (4.1),

$$\mathrm{E}(Y^t - Y^u) = 0.2 = 0.6 - 0.4 = \mathrm{E}(Y^t) - \mathrm{E}(Y^u). \tag{4.5}$$

Table 4.1: Charter school example population of potential outcomes and ATE.

Student type	Probability	Y^u	Y^t	$Y^t - Y^u$
1	0.3	0	0	0
2	0.3	0	1	1
3	0.1	1	0	-1
4	0.3	1	1	0
Mean		0.4	0.6	0.2

There are some important limitations of the ATE, including the following.

- Zero ATE does not mean zero effect (e.g., it could affect variance).
- ATE compares a universe where everybody is treated to a universe where nobody is treated, which may be unrealistic; often we are interested in more marginal policy changes.

See Section 4.5.2 of Kaplan (2022) for details and examples.

4.1.4 ATE Identification

Besides their potential outcomes, each individual has a treatment dummy X such that their observed outcome Y is

$$Y = (1 - X)Y^u + XY^t. \tag{4.6}$$

That is, if $X = 0$, then $Y = Y^u$, whereas if $X = 1$, then $Y = Y^t$.

Assumption A4.1 (SUTVA)**.** Everyone with $X = 1$ receives the same treatment, and one individual's treatment does not affect any other individual's potential outcomes.

Assumption A4.1 is usually just called SUTVA, but the main part of it is often called **no interference** (or **non-interference**).

Assumption A4.2 (independence)**.** Treatment is independent of the potential outcomes: $X \perp\!\!\!\perp (Y^u, Y^t)$.

Assumption A4.2 has many names: **independence**, **ignorability**, or **unconfoundedness**. The combination of A4.2 and A4.3 is sometimes called **strong ignorability**. For more detail, history, and discussion, see Imbens and Wooldridge (2007).

Assumption A4.3 (overlap)**.** There is strictly positive probability of both treatment and non-treatment: $0 < \mathrm{P}(X = 1) < 1$.

Assumption A4.3 is intuitive: if everybody (or nobody) is treated, then it's impossible to compare treated and untreated outcomes. For example, if $\mathrm{P}(X = 1) = 0$, then nobody is treated, so it's impossible to learn about $\mathrm{E}(Y^t)$ because Y^t is never observed. Although trivial in this simple context, overlap becomes more important to consider when other conditioning variables are included.

Theorem 4.1 formally states the ATE identification result. Intuitively, the key is that A4.2 allows us to observe representative samples of both Y^u and Y^t; treatment cannot be chosen or assigned based on an individual's potential outcomes. Mathematically, A4.2 implies that the means of the potential outcomes do not statistically depend on the treatment X:

$$\mathrm{E}(Y^t) = \mathrm{E}(Y^t \mid X = 1), \quad \mathrm{E}(Y^u) = \mathrm{E}(Y^u \mid X = 0). \tag{4.7}$$

This condition is called **mean independence**: conditioning on X does not affect the mean of Y^t or of Y^u. Independence implies mean independence; mean independence is weaker than independence. We observe $Y = Y^t$ when $X = 1$ and $Y = Y^u$ when $X = 0$, so

$$\mathrm{E}(Y^t \mid X = 1) = \mathrm{E}(Y \mid X = 1), \quad \mathrm{E}(Y^u \mid X = 1) = \mathrm{E}(Y \mid X = 0). \tag{4.8}$$

Combining (4.7) and (4.8), this says that the population mean of the treated *potential* outcome, $\mathrm{E}(Y^t)$, equals the mean of the *observed* outcome in the treated population, $\mathrm{E}(Y \mid X = 1)$. Thus, $\mathrm{E}(Y^t) = \mathrm{E}(Y \mid X = 1)$ is identified. Similarly, $\mathrm{E}(Y^u) = \mathrm{E}(Y \mid X = 0)$ is identified, so $\mathrm{E}(Y^t) - \mathrm{E}(Y^u)$ is identified.

Theorem 4.1 (ATE identification)**.** *Under A4.1–A4.3, the ATE is identified:*

$$\mathrm{E}(Y^t - Y^u) = \mathrm{E}(Y^t) - \mathrm{E}(Y^u) = \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0),$$

which is also the slope in the linear CMF model $\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x$*. More generally, A4.2 can be replaced by the mean independence condition in* (4.7)*.*

Proof. Using the above,

$$\begin{aligned}
\mathrm{ATE} \equiv \overbrace{\mathrm{E}(Y^t - Y^u)}^{\text{use linearity, (4.1)}} &= \overbrace{\mathrm{E}(Y^t) - \mathrm{E}(Y^u)}^{\text{use (4.7)}} \\
&= \overbrace{\mathrm{E}(Y^t \mid X = 1)}^{\text{use (4.8)}} - \overbrace{\mathrm{E}(Y^u \mid X = 0)}^{\text{use (4.8)}} \\
&= \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0).
\end{aligned}$$

□

Beyond our scope...

We can also learn about "quantile treatment effects" like the median treatment effect, defined as the difference between the medians of the treated and untreated potential outcome distributions. The same identification argument goes through if we assume median independence instead of mean independence; with full independence, all quantile treatment effects (and the ATE) are identified. For example, see Chapter 6 of Kaplan (2021).

Discussion Question 4.1 (college and wage)**.** Let $X = 1$ if an individual has a college degree (the "treatment") and otherwise $X = 0$. Let Y be the individual's wage at age 45, with Y^u and Y^t the potential outcomes. Explain specifically why A4.2 is violated.

4.1.5 SUTVA Violations

As alluded to above, SUTVA can be violated in many ways, especially in economics. This is not about sampling, or randomization, or data; it is about the potential outcomes framework itself. Without SUTVA, it's unclear what "treatment effect" even means.

One common violation of SUTVA is from **spillover effects** that benefit even untreated individuals. That is, the treatment's benefit "spills over" into untreated individuals. Perhaps the treated individuals can share the treatment itself with others, or perhaps others benefit from the improved outcomes of treated individuals.

Example 4.4. Consider a treatment that provides treated individuals with helpful information about financial planning. Treated individuals might share such information with their untreated friends and family. Thus, an untreated individual's outcome may depend on whether or not their friend is treated. This spillover effect violates the "no interference" part of SUTVA.

Example 4.5. Consider a "treatment" that leads to less binge drinking among treated individuals. Even if the treatment itself is not shared, the reduction in binge drinking may reduce social pressure and result in less binge drinking among untreated individuals. Here, untreated individuals are affected by the treatment through the changed behavior of treated individuals. This spillover effect violates SUTVA.

Another common violation of SUTVA is from **general equilibrium effects** (Section 3.5), such as changing market prices.

Example 4.6 (Kaplan video)**.** Consider a new agricultural technology hoping to increase cacao farmers' earnings (through increased productivity). If only one farmer gets this treatment (technology), then she benefits from increased production, selling more cacao at the current global price. But if all farmers in the world get the technology, then the global cacao supply curve shifts and the price drops. Thus, each farmer's untreated and treated potential outcomes (earnings) are affected by all other farmers' treatment status, which affects the market equilibrium price. This violates SUTVA.

Example 4.7. Consider the "treatment" that provides a subsidy for buying a house. This increases demand, which increases prices. This general equilibrium effect violates SUTVA.

Discussion Question 4.2 (cash transfer spillovers)**.** Consider the effect of income on food consumption (Y) in a rural village. Consider an "unconditional cash transfer" program (like GiveDirectly) that (potentially) gives the equivalent of \$1000.00 to a treated individual. Describe different possible spillover effects that would violate SUTVA.

Beyond our scope...

Check your intuition at `https://doi.org/10.3982/ECTA17945` that reports estimates of such spillover effects.

4.2 Linear Structural Model

To identify parameters in a structural model, generally we need some sort of exogeneity condition that says X is unrelated to other causal determinants of Y. Below are some examples.

4.2.1 Fixed Coefficients

Consider the linear structural model

$$Y = \beta_0 + \beta_1 X + U, \tag{4.9}$$

where the unobserved scalar U captures the combined effect on Y of everything besides the common linear effect β_1 of X (and β_0 and β_1 are non-random parameter values). That is, U contains heterogeneity (if some individuals' effect of X is above β_1, or below), as well as nonlinearity in X (like if Y also depends on X^2), as well as omitted variables (like if some other Q has an effect on Y). If all these other effects are "unrelated" to X, then X is called **exogenous** and β_1 is identified; if not, then X is called **endogenous**. Mathematically, here "exogenous" means uncorrelated. In other contexts, "exogenous" may required mean independence or independence (which here are sufficient but not necessary).

Theorem 4.2 (linear structural identification)**.** *Given* (4.9)*, if* $\mathrm{Cov}(X, U) = 0$*, then* β_1 *is identified and equals the slope of the linear projection* $\mathrm{LP}(Y \mid 1, X)$*.*

Proof. The slope of $\mathrm{LP}(Y \mid 1, X)$ is

$$\frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)} = \frac{\mathrm{Cov}(\beta_0 + \beta_1 X + U, X)}{\mathrm{Var}(X)} = \frac{\beta_1 \mathrm{Cov}(X, X) + \mathrm{Cov}(U, X)}{\mathrm{Var}(X)} = \frac{\beta_1 \mathrm{Var}(X) + 0}{\mathrm{Var}(X)} \quad \square$$
$$= \beta_1.$$

Discussion Question 4.3 (college and wage: endogeneity)**.** Let $X = 1$ if an individual has a college degree and otherwise $X = 0$. Let Y be the individual's wage at age 45. Explain one real-world reason why $\mathrm{Cov}(X, U) \neq 0$.

4.2.2 Random Coefficients

Consider the linear structural **random coefficients** model

$$Y = U_0 + U_1 X, \tag{4.10}$$

where U_0 and U_1 are unobserved random variables. That is, each individual is represented by (Y, U_0, U_1, X); you can think of "random" as just meaning "individual-specific." This model more explicitly shows the heterogeneity in the intercept and slope. If X is binary, then the potential outcomes model in (4.6) can be rewritten as

$$Y = Y^u + (Y^t - Y^u)X, \tag{4.11}$$

which is (4.10) with $U_0 = Y^u$ and $U_1 = Y^t - Y^u$ (the individual's treatment effect).

Discussion Question 4.4 (college and wage: random coefficients)**.** Let $X = 1$ if an individual has a college degree and otherwise $X = 0$. Let Y be the individual's wage at age 45.

a) How do you interpret the economic meaning of U_0?
b) How do you interpret the economic meaning of U_1?
c) Why do you think individuals have different U_0?
d) Why do you think individuals have different U_1?

Theorem 4.3 (linear random coefficient identification)**.** *Given* (4.10)*, if* U_0 *and* U_1 *are mean-independent of* X*, then* $\mathrm{E}(U_0)$ *and* $\mathrm{E}(U_1)$ *are identified and equal to the linear CMF intercept and slope, respectively.*

Proof. Take the conditional mean of (4.10):

$$\mathrm{E}(Y \mid X) = \mathrm{E}(U_0 + U_1 X \mid X) = \mathrm{E}(U_0 \mid X) + \mathrm{E}(U_1 \mid X)X = \mathrm{E}(U_0) + \mathrm{E}(U_1)X$$

by mean independence. That is, the conditional mean of Y given X is linear (affine) in X, with non-random intercept and slope $\mathrm{E}(U_0)$ and $\mathrm{E}(U_1)$, respectively. □

We can connect the random coefficients model to the fixed coefficients model in (4.9). Rewrite (4.10) as

$$\begin{aligned} Y &= U_0 + U_1 X \\ &= U_0 + U_1 X + \overbrace{\mathrm{E}(U_0) - \mathrm{E}(U_0)}^{=0} + \overbrace{[\mathrm{E}(U_1) - \mathrm{E}(U_1)]X}^{=0} \\ &= \overbrace{\mathrm{E}(U_0)}^{\beta_0} + \overbrace{\mathrm{E}(U_1)}^{\beta_1} X + \overbrace{U_0 - \mathrm{E}(U_0) + [U_1 - \mathrm{E}(U_1)]X}^{U} . \end{aligned}$$

4.3 Nonseparable Structural Model

Consider the **all-causes model**

$$Y = h(X, \boldsymbol{U}) \tag{4.12}$$

that shows how Y is fully determined by observable scalar X and unobserved vector $\boldsymbol{U}$, through the function $h(\cdot)$. That is, $\boldsymbol{U}$ contains all determinants of Y besides X, and thus is very large. This model is called **nonseparable** because the X and unobservables enter $h(\cdot)$ together, not additively separable like $Y = f(X) + g(\boldsymbol{U})$. The nonseparable model is more general because it still allows for additive separability but does not require it.

The **average structural function** (ASF) is a common object of interest, defined as

$$\mathrm{ASF}(x) \equiv \mathrm{E}[h(x, \boldsymbol{U})], \tag{4.13}$$

where the expectation is with respect to the unconditional distribution of $\boldsymbol{U}$. Like the CMF, the ASF is a non-random function (because it plugs in a non-random x and then averages out the $\boldsymbol{U}$).

The **average structural effect** (ASE) is the partial derivative of the ASF:

$$\mathrm{ASE}(x) \equiv \frac{\partial}{\partial x}\mathrm{ASF}(x) = \mathrm{E}[\frac{\partial}{\partial x}h(x, \boldsymbol{U})]. \tag{4.14}$$

If x is discrete instead of continuous, then as usual the partial derivative can be replaced by a discrete difference like

$$\mathrm{ASE} = \mathrm{ASF}(1) - \mathrm{ASF}(0) = \mathrm{E}[h(1, \boldsymbol{U}) - h(0, \boldsymbol{U})].$$

As with the ATE, due to linearity of expectation, we can either think of the ASE as the difference between two points on the ASF (or derivative), or the mean of the individual-level causal effects. For example, in the binary X case, the causal effect of changing $X = 0$ to $X = 1$ is

$$C(\boldsymbol{U}) \equiv h(1, \boldsymbol{U}) - h(0, \boldsymbol{U}), \tag{4.15}$$

which depends on an individual's $\boldsymbol{U}$ (some individuals' Y may be more responsive to X changes than others'). The average such causal effect is

$$\mathrm{E}[C(\boldsymbol{U})] = \mathrm{E}[h(1, \boldsymbol{U}) - h(0, \boldsymbol{U})] = \mathrm{E}[h(1, \boldsymbol{U})] - \mathrm{E}[h(0, \boldsymbol{U})] = \mathrm{ASF}(1) - \mathrm{ASF}(0) = \mathrm{ASE}\,. \tag{4.16}$$

To connect back with the ATE, first consider a binary X. An individual with unobserved $\boldsymbol{U}$ has potential outcomes

$$Y^u = h(0, \boldsymbol{U}), \quad Y^t = h(1, \boldsymbol{U}). \tag{4.17}$$

That is, the model says if we change the individual's $X = 0$ to $X = 1$, their Y will change from $h(0, \boldsymbol{U})$ to $h(1, \boldsymbol{U})$.

Theorem 4.4. *Given the structural all-causes model in (4.12) with binary $X \in \{0, 1\}$ and the potential outcomes in (4.17), if $X \perp\!\!\!\perp \boldsymbol{U}$, then the slope coefficient β_1 of the CMF $\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x$ equals the ATE and equals the ASF.*

Proof. Given $X \perp\!\!\!\perp \boldsymbol{U}$, (4.12), and (4.13),

$$\mathrm{E}(Y \mid X = 1) = \mathrm{E}[h(X, \boldsymbol{U}) \mid X = 1] = \mathrm{E}[h(1, \boldsymbol{U}) \mid X = 1] = \mathrm{E}[h(1, \boldsymbol{U})] = \mathrm{ASF}(1),$$

and similarly

$$\mathrm{E}(Y \mid X = 0) = \mathrm{E}[h(0, \boldsymbol{U})] = \mathrm{ASF}(0).$$

The CMF slope equals $\mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)$, which thus equals $\mathrm{ASF}(1) - \mathrm{ASF}(0)$, which is the average structural effect of X on Y. Taking expectations of (4.17), $\mathrm{E}(Y^u) = \mathrm{ASF}(0)$ and $\mathrm{E}(Y^t) = \mathrm{ASF}(1)$, so additionally the ASE equals $\mathrm{E}(Y^t) - \mathrm{E}(Y^u)$, which equals the ATE. □

Theorem 4.5. *Given the structural all-causes model in (4.12), if $X \perp\!\!\!\perp \boldsymbol{U}$, then the ASF is identified and equals the CMF (and thus the ASE is the derivative or difference of the CMF).*

Proof. For any x,

$$\mathrm{E}(Y \mid X = x) = \mathrm{E}[h(X, \boldsymbol{U}) \mid X = x] = \mathrm{E}[h(x, \boldsymbol{U}) \mid X = x] = \mathrm{E}[h(x, \boldsymbol{U})] = \mathrm{ASF}(x),$$

where the first equality is from plugging in (4.12), the second is because it conditions on $X = x$, the third is from $X \perp\!\!\!\perp \boldsymbol{U}$, and the fourth is the definition of ASF in (4.13). □

Note: although the nonseparable model may seem fancy, it's still essentially asking: what if we changed every single individual in the population from $X = x$ to $X = x + 1$? (Or $X = x + \mathrm{d}x$.) Often a policy only affects certain individuals' X values (and such "marginal" individuals may differ in important ways from the population as a whole).

Discussion Question 4.5 (wage and education: nonseparable)**.** Imagine an "audit study" where fake resumes are posted to apply to jobs online, and years of experience $X \in \{0, 1, 2, \ldots, 15\}$ is randomized while holding other applicant characteristics constant (or varying them independently of X). Let $Y = 1$ if the employer requests a follow-up interview, otherwise $Y = 0$.

a) Interpret the identification result in Theorem 4.5 in terms of this example.
b) How/does the identification result help us estimate the effect of interest from our audit study data? Explain.

Chapter 5

Identification by Conditional Independence

Unit learning objectives for this chapter

5.1. Explain how the intuition of causal identification from independence extends to a conditional setting. [TLO 2]

The intuition for identification by independence can be extended to conditional independence, for both treatment effects and structural models.

5.1 Conditional Average Treatment Effect

Consider the ATE for the subpopulation of individuals with $\boldsymbol{W} = \boldsymbol{w}$. For example, this could be the subpopulation of individuals who are 40 years old, have 16 years of education, and are married. This is called the **conditional average treatment effect** (CATE), here denoted

$$\text{CATE}(\boldsymbol{w}) \equiv \mathrm{E}(Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w}) = \mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}) - \mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}), \tag{5.1}$$

where as in (4.1) the equality is due to the linearity of the expectation operator.

By the law of iterated expectations, the ATE can be written as

$$\text{ATE} = \mathrm{E}[\text{CATE}(\boldsymbol{W})], \tag{5.2}$$

where the expectation is with respect to the population distribution of $\boldsymbol{W}$.

Imagine we run an experiment where we randomize treatment, but the treatment probability is higher for unemployed individuals, and the outcome Y is wage one year later. Let $X = 1$ if treated and $X = 0$ if untreated; let $W = 1$ if unemployed and

$W = 0$ otherwise. We randomize with $\mathrm{P}(X = 1 \mid W = 1) = 0.8$ and $\mathrm{P}(X = 1 \mid W = 0) = 0.1$. Our earlier independence assumption is likely violated. For simplicity, imagine the treatment is useless, so $Y^t = Y^u = Y$ for everyone. Assuming the unemployed individuals tend to have lower wages, then a simple comparison of treated and untreated wages misleadingly suggests the treatment has a negative effect. For example, simplifying further, imagine all unemployed individuals have $Y = 15$ and all employed individuals have $Y = 25$, and there are 10 unemployed individuals and 20 employed individuals. Given the treatment probabilities, the treatment group consists of 8 of the 10 unemployed individuals with $Y = 15$, plus 2 of the 20 employed individuals with $Y = 25$; altogether, 10 individuals, with average wage 17. The remaining individuals are untreated, with average wage $[(2)(15) + (18)(25)]/(2 + 18) = 24$, much higher than the treated group! This is because the independence assumption A4.2 fails: X is not independent of (Y^u, Y^t).

However, we can still identify the true ATE by using conditional independence. Given W, X is independent of potential outcomes. For example, if we only look at unemployed individuals, then we have a randomized experiment where A4.2 holds; and similarly if we only look at employed individuals. Thus, the conditional ATEs are identified, and the ATE is identified by taking a weighted average of the CATEs (rather than pooling the data like above).

Assumption A5.1 (conditional independence)**.** Treatment X is conditionally (on $\boldsymbol{W}$) independent of the potential outcomes: $(Y^u, Y^t) \perp\!\!\!\perp X \mid \boldsymbol{W}$.

Assumption A5.2 (overlap)**.** There is strictly positive probability of both treatment and non-treatment for every subpopulation: $0 < \mathrm{P}(X = 1 \mid \boldsymbol{W} = \boldsymbol{w}) < 1$ for all $\boldsymbol{w}$.

The key argument again relies on the (conditional) independence assumption. Assumption A5.1 implies

$$\begin{aligned} \mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}) &= \mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) \\ \mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}) &= \mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 0). \end{aligned} \tag{5.3}$$

This condition is called **conditional mean independence**: after conditioning on $\boldsymbol{W} = \boldsymbol{w}$, further conditioning on X does not affect the conditional mean of Y^t or Y^u. Given (4.6),

$$\begin{aligned} \mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) &= \mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) \\ \mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 0) &= \mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 0). \end{aligned} \tag{5.4}$$

This is an example of **nonparametric identification** because we have not restricted the CMF $\mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = x)$ to be linear or quadratic or have any other specific functional form (parameterization). Even if we end up estimating the CMF parametrically, it is still reassuring that our underlying identification argument does not rely on our knowing the true functional form.

Theorem 5.1 (CATE identification)**.** *Under A4.1, A5.1, and A5.2, each CATE is identified:*

$$\mathrm{CATE}(\boldsymbol{w}) = \mathrm{E}[Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w}] = \mathrm{E}(Y \mid X = 1, \boldsymbol{W} = \boldsymbol{w}) - \mathrm{E}(Y \mid X = 0, \boldsymbol{W} = \boldsymbol{w}).$$

Thus, the ATE is also identified. More generally, A5.1 can be replaced by conditional mean independence as in (5.3)*.*

Proof. Using the above ingredients,

$$\begin{aligned}
\mathrm{CATE}(\boldsymbol{w}) &\equiv \overbrace{\mathrm{E}(Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w})}^{\text{use linearity, (5.1)}} \\
&= \overbrace{\mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}) - \mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w})}^{\text{use (5.3)}} \\
&= \overbrace{\mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}, X = 1)}^{\text{use (5.4)}} - \overbrace{\mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 0)}^{\text{use (5.4)}} \\
&= \mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) - \mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 0),
\end{aligned}$$

which is a feature of (only) the joint population distribution of observables $(Y, \boldsymbol{W}, X)$. By (5.2), the ATE is thus also identified. ☐

Beyond our scope...

How can we estimate the CATE using the identification result in Theorem 5.1? In principle, given iid data (or otherwise restricted dependence), we can consistently estimate any feature of the joint distribution of $(Y, X, \boldsymbol{W})$. For example, if X and W are both binary, then $\mathrm{E}(Y \mid W = 0, X = 0)$ can be estimated by $\widehat{\mathrm{E}}(Y \mid W = 0, X = 0)$, the sample mean of the Y_i for observations with $W_i = 0$ and $X_i = 0$. However, if W is continuous, then $\mathrm{P}(W_i = w) = 0$ for any $w \in \mathbb{R}$, so this estimation approach fails. We need to either assume the CMF is linear or quadratic (or some other specific functional form), or else use nonparametric regression, as introduced in Part V of Kaplan (2021).

Discussion Question 5.1 (doctor certification)**.** Let $W \in \mathbb{R}$ be a continuous scalar random variable representing a doctor's quality, and let $X = \mathbb{1}\{W \geq 0\}$ be a dummy variable for whether the doctor receives a publicly visible certification as being high-quality. Let $Y^t \in [0, 10]$ be the doctor's patient satisfaction rating in the world where the doctor is certified, and $Y^u \in [0, 10]$ the rating in the world where the doctor is not certified (but everything else is identical, including true quality W).

a) Give a specific, real-world reason why probably $(Y^u, Y^t) \perp\!\!\!\perp X$ fails; explain both intuitively and mathematically. (If it helps: try graphing the functions $\mu_t(w) = \mathrm{E}(Y^t \mid W = w)$ and $\mu_u(w) = \mathrm{E}(Y^u \mid W = w)$.)

b) Explain how it is possible to satisfy Assumption A5.1 even if patient satisfaction increases with true quality W.
c) Explain why Assumption A5.2 (overlap) fails, and intuitively why thus we cannot estimate $\mathrm{E}(Y \mid W = w, X = 1) - \mathrm{E}(Y \mid W = w, X = 0)$.

Beyond our scope...

The overlap problem in DQ 5.1 can be addressed by (roughly speaking) comparing doctors who are just barely above zero to doctors who are just barely below: their W is very similar, but the former have $X = 1$ while the latter have $X = 0$. This approach is called **regression discontinuity** and is covered in the ECON 9446/9447 sequence.

5.2 Linear Structural Model

Consider the linear structural model

$$Y = \beta_0 + X\beta_1 + \boldsymbol{W}'\boldsymbol{\beta}_2 + U. \tag{5.5}$$

XXX

XXX

5.3 Nonseparable Structural Model

Extending (4.12), consider the nonseparable all-causes model

$$Y = h(X, \boldsymbol{W}, \boldsymbol{U}), \tag{5.6}$$

where $(Y, X, \boldsymbol{W})$ is observable but not $\boldsymbol{U}$.

First consider binary X. For an individual with $(\boldsymbol{w}, \boldsymbol{u})$, the causal effect (structural effect) of X on Y is

$$C(\boldsymbol{w}, \boldsymbol{u}) \equiv h(1, \boldsymbol{w}, \boldsymbol{u}) - h(0, \boldsymbol{w}, \boldsymbol{u}). \tag{5.7}$$

If we condition on $\boldsymbol{w}$ but average out $\boldsymbol{u}$,

$$\mathrm{CASE}(\boldsymbol{w}) = \mathrm{E}[C(\boldsymbol{w}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}], \tag{5.8}$$

where the expectation is taken with respect to the conditional distribution of $\boldsymbol{U}$ given $\boldsymbol{W} = \boldsymbol{w}$. XXX CONNECT w/ potential outcomes from CATE

XXX CASE (general) XXX

Chapter 6

OVB and Proxy Variables

Unit learning objectives for this chapter

6.1. Define terms and concepts related to omitted variable bias and proxy variables. [TLO 1]

6.2. Describe how proxy variables can help reduce omitted variable bias, both intuitively and mathematically. [TLO 3]

This chapter first describes and quantifies the problem known as omitted variable bias, for linear structural models. Then, it shows how "proxy variables" can help reduce this bias.

This problem only really applies to estimating causal effects. For description, for example, if we want to estimate the linear projection slope of $\mathrm{LP}(Y \mid 1, X)$, then it doesn't matter what other variables there are; the linear projection depends only on Y and X. (Mathematically, you could ask about bias in the estimated coefficient on X in $\mathrm{LP}(Y \mid 1, X, Q)$ if Q is omitted and instead $\mathrm{LP}(Y \mid 1, X)$ is estimated, but that's not usually a situation faced in practice.) For prediction, we do not care about coefficients, only prediction accuracy; omitting a predictor may make our accuracy worse, but we wouldn't say it's "biased."

6.1 Omitted Variable Bias

XXX

XXX REVISE! Linear structural model:

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + Q\gamma + V. \tag{6.1}$$

Assume if observe Q then can estimate by OLS, i.e., is LP model in error form, i.e., $\mathrm{E}[(\boldsymbol{X}', Q)V] = \boldsymbol{0}'$. If Q not observed, then with $U \equiv Q\gamma + V$,

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + U, \tag{6.2}$$

but if Q and $\boldsymbol{X}$ are related, then U is not LP error. Let

$$\mathrm{LP}(Q \mid \boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\delta}, \quad R \equiv Q - \boldsymbol{X}'\boldsymbol{\delta}. \tag{6.3}$$

Plug into (6.1),

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{\delta} + R)\gamma + V = \boldsymbol{X}'(\boldsymbol{\beta} + \gamma\boldsymbol{\delta}) + (V + R\gamma). \tag{6.4}$$

Note $V + R\gamma$ satisfies the LP error property ..... Thus, OLS is consistent for $\boldsymbol{\beta} + \gamma\boldsymbol{\delta}$, meaning the asymptotic bias is $\gamma\boldsymbol{\delta}$.

Note if all $\delta_j = 0$ except $\delta_k \neq 0$, then $\delta_k = \mathrm{Cov}(Q, X_k)/\mathrm{Var}(X_k)$, which is easier to think about than a general partial correlation (LP coefficient).

Discussion Question 6.1 (wage OVB)**.** Let Y be log wage, X_1 experience, X_2 years of education, and Q unobserved "ability," and assume that the structural model $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + Q\gamma + V$ has structural error term V satisfy $\mathrm{E}[(1, X_1, X_2, Q)V] = \mathbf{0}'$. Also assume for simplicity $\mathrm{LP}(Q \mid 1, X_1, X_2) = \delta_0 + X_1\delta_1 + X_2\delta_2$ has $\delta_1 = 0$.

a) Do you think $\delta_2 < 0$, $\delta_2 > 0$, or $\delta_2 = 0$? Explain why.
b) Do you think β_2 is < 0, > 0, or $= 0$? Explain.
c) Do you think γ is < 0, > 0, or $= 0$? Explain.
d) Given the above, would $\operatorname{plim}_{n\to\infty} \hat{\beta}_2$ be $< \beta_2$, $> \beta_2$, or $= \beta_2$? Explain.

XXX

6.2 Proxy Variables

XXX

6.2.1 Perfect Proxy

XXX ASSUMPTIONS

XXX REVISE Wooldridge: "redundant" proxy Z satisfies

$$\mathrm{E}(Y \mid \boldsymbol{X}, Q, Z) = \mathrm{E}(Y \mid \boldsymbol{X}, Q), \tag{6.5}$$

so that $\mathrm{E}(V \mid \boldsymbol{X}, Q, Z) = 0$.

$$\mathrm{LP}(Q \mid \boldsymbol{X}, Z) = \boldsymbol{X}'\boldsymbol{\rho} + Z\theta_1. \tag{6.6}$$

Plugging in,

$$Y = \boldsymbol{X}'(\boldsymbol{\beta} + \gamma\boldsymbol{\rho}) + \gamma\theta_1 Z + (\gamma R + V) \tag{6.7}$$

where

$$R \equiv Q - \mathrm{LP}(Q \mid \boldsymbol{X}, Z). \tag{6.8}$$

2nd assumption is for PERFECT proxy:

$$\mathrm{LP}(Q \mid \boldsymbol{X}, Z) = \mathrm{LP}(Q \mid 1, Z) = \theta_0 + Z\theta_1. \tag{6.9}$$

Plug in:

$$Y = (\beta_0 + \gamma\theta_0) + \boldsymbol{X}'_{(-1)}\boldsymbol{\beta}_{(-1)} + \gamma\theta_1 Z + (\gamma R + V) \tag{6.10}$$

where

$$R \equiv Q - \mathrm{LP}(Q \mid \boldsymbol{X}, Z). \tag{6.11}$$

XXX THEOREM

6.2.2 Imperfect Proxy

XXX ASSUMPTIONS

XXX THEOREM

Discussion Question 6.2 (recidivism and therapy)**.** Consider the causal effect of a particular cognitive behavioral therapy (CBT) program on the future criminal activity of current prison inmates. Specifically, there is a group of individuals who were previously in prison, with $X = 1$ if they participated in CBT and $X = 0$ if not, and who were then tracked for five years after their release from prison, with Y the number of additional days spent in prison during that five-year window.

a) Prisoners are more likely to be assigned to CBT if they committed a more severe crime; prisoners are also more likely to commit future crimes (and more severe future crimes) if their initial crime is more severe. Explain which direction of OVB this generates.
b) We also observe Z, the length (in days) of the prisoner's initial prison sentence. Explain with words and equations the conditions under which Z would be a perfect proxy

However, if a variable is a really bad proxy, then it can actually worsen OVB, as well as increasing standard errors. Thus, you need to think carefully about whether a variable actually should proxy for a particular omitted variable, rather than simply including all available variables in the data. For example, see the simple example at the very end of Chapter 4 (page 72) of Wooldridge (2010).

Exercises

Exercise I.1. Write a Stata do-file as follows. In general, each step corresponds to one line of code, except where otherwise noted. The data files are available at:
https://drive.google.com/file/d/0B-_LUSJVBv20SjBYd2pwYkYtcnc/view?resourcekey=0-DMCuTq__SV1c0PxaQ-wTOQ
https://drive.google.com/file/d/0B-_LUSJVBv20U2E2R2tBWnItaU0/view?resourcekey=0-dkguDqOtIoH5VWJgm-4T4g

a. Include the usual top-of-file items:

   i. Make the first line a "comment" (starting with an asterisk) with your name, the class name, and today's date.
   ii. Clear all variables in memory with `clear all`
   iii. Close any log file that may currently be open, without displaying an error if none is open, with `capture log close`
   iv. Issue the command `set more off` so that Stata doesn't wait for your input if there's more than one screen of output.
   v. Change the current directory to the one where you have downloaded the raw data and have saved this do-file, using the `cd` command.
   vi. Start writing a plaintext log to a file with suffix ".log", replacing the existing file if applicable (with the `replace` option).

b. Read into memory the data in file `Kaplan_Stata1_fake_data_grades.csv` using the command `insheet`:
   `insheet using "Kaplan_Stata1_fake_data_grades.csv" , clear`

c. Using a `keep if` statement, keep only rows for undergraduates, who are identified by their student type being `UG`.

d. Create a new variable named `cl_grade_num` that translates the string variable `cl_grade` into the corresponding numeric values. For example, if row 7 contains `cl_grade` equal to `D`, then the new variable should equal one; A=4, B=3, C=2, D=1, F=0. Note: this step requires multiple lines of code; the first is a `generate` command, and the rest are `replace` commands.

e. Create (with command `generate`) another new variable: name it `cl_grade_pts`, and store the product of `cl_units` and `cl_grade_num`.

f. Collapse (with command `collapse`) the data to one row per student, calculating the sum of `cl_units` and `cl_grade_pts`.

g. Create a new variable named `s_GPA` as the quotient of the summed `cl_grade_pts` and the summed `cl_units`; this is the grade point average (GPA), the average of the grades weighted by the units per class.

h. Drop (with `drop`) the variables containing the summed `cl_grade_pts` and summed `cl_units`.

i. Sort by `s_id`.

j. Check whether `s_id` is a unique identifier: `isid s_id`.

k. Save the dataset to a new .dta file, replacing the existing version (if applicable): `save "Kaplan_Stata1_fake_data_GPA.dta" , replace`

l. Load the data in `Kaplan_Stata1_fake_data_parents.csv` using `insheet`.

m. Rename (command `rename`) the variable `student_id` to `s_id` to match the other file's convention.

n. Convert the variable `s_id` from string to numeric, ignoring the leading `A` in each: `destring s_id, replace ignore("A")`

o. Reshape the data to have only one row per student, with variable `p_edu1` containing parent 1's education and `p_edu2` for parent 2: `reshape wide p_edu,i(s_id) j(parent)`

p. Make a new variable named `p_edu_max` that is the maximum of all variables with prefix `p_edu`, using the `egen` command with `rowmax` (ignoring missing values): `egen p_edu_max = rowmax(p_edu*)`

q. Sort by `s_id`

r. Check that `s_id` is a unique identifier with `isid`

s. Merge the data currently in memory 1:1 by `s_id` with the temporary file with GPA that you saved earlier.

t. Drop observations containing data only from the parent dataset and not from the GPA dataset, i.e., when the generated variable `_merge` equals one.

u. Order the columns in the dataset so that `s_id` is first, then `s_GPA`, then other variables: `order s_id s_GPA`

v. Print the dataset to the console/log using the `list` command (with no arguments or options). (Note: in reality, you would rarely print an entire dataset since they are usually much bigger than this artifical example.)

w. Save your dataset to a new .dta file.

x. Close your log file.

After thoroughly debugging, run your file all the way through all at once, and submit your .log file and .do file electronically through Canvas.

Exercise I.2. Write a Stata do-file as follows. The data are from a New York Times article on December 28, 1994.

a. Do the usual top-of-file items from Exercise I.1(a).
b. Run `ssc install bcuse` to ensure command `bcuse` is installed, and then load the dataset with `bcuse wine, clear`
c. View basic dataset info with Stata command `describe`
d. View the first few rows of the dataset with Stata command `list if _n<=5`
e. Rename the `alcohol` column, which measures liters of alcohol from wine (consumed per capita per year): `rename alcohol wine`
f. Add a column named `id` whose value is just 1, 2, 3, 4, 5, etc.: `generate id = _n`
g. Display the countries with fewer than 100 heart disease deaths per 100,000 people: `list country if heart<100`
h. Display the rows for the countries with the 5 lowest death rates, sorted by death rate: `sort deaths` followed by (next line) `list if _n<=5`
i. Add a column with the sum of heart and liver disease deaths per 100,000: `generate heart_plus_liver = heart + liver`
j. Generate a variable with the squared death rate: `gen deaths_sq = deaths^2`
k. Display the sorted death rates: `sort deaths` followed by `list deaths`
l. Add a column with the proportion of heart deaths to total deaths with command `generate heart_prop = heart / deaths`
m. Create a histogram of liver deaths: `histogram liver`
n. Create a scatterplot of liver death rates (vertical axis) against wine consumption (horizontal axis): `scatter liver wine`

Exercise I.3. Consider the effect of being assigned to a job training program, where assignment was randomized. The specific program was the National Supported Work Demonstration in the 1970s in the U.S. Data are originally from LaLonde (1986), via Wooldridge (2020). You will look at effects on earnings. The `train` variable indicates (randomized) assignment to job training if it equals 1, and it equals 0 otherwise.

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`
b. Load the data: `bcuse jtrain2 , clear`
c. Run `describe re78 train` and read the variable labels to understand the meaning and units of measure.

d. Run `ttest re78 , by(train) unequal` and explain in words briefly (1 sentence) what that code does.

e. Run `reg re78 train , vce(robust)` and explain in words briefly (1 sentence) what that code does.

f. Rounding to three significant figures (and including units of measure), report the estimated average effect of being assigned to training, and discuss the estimate's economic significance (magnitude).

g. Rounding to three significant figures (and including units of measure), report the corresponding 95% confidence interval, and discuss what this tells us about uncertainty (be precise).

h. Describe the "potential outcomes" in this example, and explain why the average treatment effect of assignment to job training seems to be identified.

i. If this job training program were scaled up and offered to every individual in the country, would you guess the average effect would be higher or lower (due to general equilibrium effects)? Explain in 1–2 sentences.

Exercise I.4. The data are originally from Card (1995), with individual-level observations of (log) wages, years of education, and other variables. Note the dataset lacks variable labels, but they can be found online.[1]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse card , clear`

c. Create a dummy to compare high-school (12 years education) and college (16 years education):

```
gen d_coll = .
replace d_coll=0 if educ==12
replace d_coll=1 if educ==16
```

d. Regress log wage on years of education `reg lwage educ , vce(robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction.

e. Run `reg lwage d_coll , vce(robust)` and re-phrase your above concern (about OVB) in terms of why the average treatment effect is not identified (make sure to define the potential outcomes first).

f. Run `reg lwage educ IQ , vce(robust)`

  i. Explain the conditions under which IQ would be a perfect proxy for unobserved "ability."

[1] http://fmwww.bc.edu/ec-p/data/wooldridge/card.des

ii. Briefly describe one type of "ability" that IQ does not capture (so is not a perfect proxy). Given this, do you think it's better or worse (or neither) to use IQ as a proxy for ability?
iii. Does the estimated slope change in the direction that suggests reduced OVB? Explain briefly.
iv. Discuss the economic significance of the estimated slope on `educ`.
v. Explain what the confidence interval tells us about our uncertainty; be precise and explicit about whichever population value(s) you refer to, and about sources of uncertainty, etc.

g. Run `reg lwage educ IQ exper expersq black smsa south , vce(robust)` and then briefly compare with previous results, focusing on the returns to education.

Exercise I.5. Go through the analysis in I.4 but with the `nls80` dataset, noting that now `iq` is lowercase.

Exercise I.6. Consider the causal effect of being an athlete on a college student's grades (GPA). Note the dataset lacks variable labels, but they can be found online.[2]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`
b. Load the data: `bcuse gpa2 , clear`
c. Regress GPA on the athlete dummy: `reg colgpa athlete , vce(robust)`

i. Interpret the estimated coefficient on `athlete` in terms of a conditional mean model.
ii. In terms of structural model $Y = \beta_0 + \beta_1 X + U$ (Y is GPA, X is the athlete dummy, U is the combined effect of unobserved determinants of Y), explain one reason why β_1 is not identified, and in which direction there is omitted variable bias. (Feel free to "cheat" and do the next parts first to get an idea!)
iii. Repeat your argument about identification failure, but in terms of a potential outcomes model and treatment effect.

d. Run `reg colgpa athlete female , vce(robust)` and explain why this seems to help (slightly) the omitted variable bias; try `tab athlete female` too.
e. Run `reg colgpa athlete female sat , vce(robust)` and explain what `sat` helps proxy for and why this helps reduce omitted variable bias.
f. Run `reg colgpa athlete female sat verbmath hsperc hsize hsizesq black white , vce(robust)`

i. Discuss the economic significance of the estimated coefficient on `athlete` and briefly compare with the original estimate from the simple regression in part (c).

[2] http://fmwww.bc.edu/ec-p/data/wooldridge/gpa2.des

ii. Explain what the confidence interval tells us about our uncertainty; be precise and explicit about whichever population value(s) you refer to, and about sources of uncertainty, etc.

Exercise I.7. Consider the causal effect of using a 401(k) retirement plan on net total financial assets. Variable descriptions are included in the dataset's variable labels.

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`
b. Load the data: `bcuse 401ksubs , clear`
c. Regress (net total financial) assets on the 401(k) participation dummy: `reg nettfa p401k , vce(robust)`
   i. Interpret the estimated coefficient on `p401k` in terms of a conditional mean model.
   ii. In terms of structural model $Y = \beta_0 + \beta_1 X + U$ (Y is assets, X is the 401(k) dummy, U is the combined effect of unobserved determinants of Y), explain one reason why β_1 is not identified, and in which direction there is omitted variable bias. (Feel free to "cheat" and do the next parts first to get an idea!)
   iii. Repeat your argument about identification failure, but in terms of a potential outcomes model and treatment effect.
d. Run `reg nettfa p401k inc , vce(robust)` and explain why this seems to help the omitted variable bias; try `reg p401k inc` too.
e. Run `reg nettfa p401k inc marr male age fsize , vce(robust)`
   i. From the potential outcomes perspective: what is the name and interpretation of the population object we hope to estimate by the coefficient on `p401k`?
   ii. Discuss the economic significance of the estimated coefficient on `p401k` and briefly compare with the original estimate from the simple regression in part (c).
   iii. Explain what the confidence interval tells us about our uncertainty; be precise and explicit about whichever population value(s) you refer to, and about sources of uncertainty, etc.

Exercise I.8. Consider the relationship between an infant's birthweight (which when too low is associated with other negative health outcomes) and the mother's cigarette smoking. Note the dataset lacks variable labels, but they can be found online.[3]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`
b. Load the data: `bcuse bwght , clear`
c. Create a dummy to compare no smoking to any smoking: `gen d_smk = (cigs>0)`

[3] http://fmwww.bc.edu/ec-p/data/wooldridge/bwght.des

d. Regress log birthweight on the amount of smoking `reg lbwght cigs , vce( robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction. (Feel free to "cheat" and look below to get ideas.)

e. Run `reg lbwght d_smk , vce(robust)` and re-phrase your above concern (about OVB) in terms of why the average treatment effect is not identified (make sure to define the potential outcomes first).

f. Run `reg lbwght cigs motheduc , vce(robust)`

   i. Explain mathematically how OVB can be reduced by using `motheduc` as a proxy for unobserved mother's knowledge about prenatal health, even if it is not a perfect proxy.

   ii. Does the estimated slope change (when adding `motheduc` as a control variable) in the direction that suggests reduced OVB? Explain briefly.

   iii. Discuss the economic significance of the estimated slope on `cigs`.

   iv. Explain what the confidence interval tells us about our uncertainty; be precise and explicit about whichever population value(s) you refer to, and about sources of uncertainty, etc.

g. Provide a reason/argument why even conditional on `motheduc`, `d_smk` is not (mean) independent of the potential outcomes.

Part II

Instrumental Variables

Chapter 7
Instrumental Variables

Unit learning objectives for this chapter

7.1. XXX [TLO 1]

This chapter XXX

7.1 XXX

XXX

Discussion Question 7.1 (XXX)**.** XXX

a) XXX
b) XXX
c) XXX
d) XXX

Chapter 8

Generalized Method of Moments

Unit learning objectives for this chapter

8.1. XXX [TLO 1]

This chapter XXX

8.1 XXX

XXX

Discussion Question 8.1 (XXX). XXX

a) XXX
b) XXX
c) XXX
d) XXX

Exercises

In Stata, the `ivreg2` command (available on SSC) helps automatically run weak identification and over-identification tests, with the help of `ranktest`, as well as GMM estimators (and yet other estimators); you can install both from SSC with:

```
ssc install ranktest
ssc install ivreg2
```

Exercise II.1. This exercise looks at the impact of participation in a 401(k) retirement plan (dummy variable `p401k`) on an individual's net total financial assets (`nettfa`), using 401(k) eligibility (dummy variable `e401k`) as an instrument.

a. For this example, describe an individual's potential outcomes.

b. For this example, describe who is a "complier" and who is a "never-taker."

c. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

d. Load the data (and look at variable labels to see descriptions and units of measure):
   `bcuse 401ksubs , clear`

e. Regress net total financial assets on 401(k) participation
   `reg nettfa p401k , vce(robust)`
   and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction.

f. Regress net total financial assets on 401(k) eligibility:
   `reg nettfa e401k , vce(robust)` and interpret the estimated coefficient on `e401k`

g. Explain what it would mean for 401(k) eligibility to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

h. Compute the IV estimator, CI, and corresponding tests:
   `ivreg2 nettfa (p401k = e401k) , robust`

   i. Describe the IV estimand in this example.
   ii. Discuss the economic significance of the estimate.

iii. Explain what the confidence interval tells us about our uncertainty about the true population value; be precise and explicit.

iv. Explain why you are or are not worried about weak instruments in this case; refer to specific Stata output.

v. Explain what the J-test results suggest about the model.

i. Run

```
ivreg2 nettfa (p401k=e401k) , robust gmm2s center
gmm (nettfa - p401k*{p401k} - {_cons}) , instruments(e401k) nolog
   vce(robust) twostep
```

and briefly compare with the estimate/CI from part (h).

j. Run `ivreg2 nettfa (p401k = e401k) inc age marr fsize , robust` and say briefly if the change (compared to above without control variables) in the estimated effect is economically significant, as well as if/how this changes our uncertainty about the true population value.

Exercise II.2. The following analyzes data originally from Graddy (1995). The goal is to estimate the demand curve for a particular type of fish (whiting) in a particular (large) fish market in New York City. Prices and quantities are in logs (so the slope is approximately an elasticity); specifically, the (average) daily price was measured in dollars per pound of fish, and the daily quantity in pounds sold, and the natural log was taken of each. The weather is used as the (hopefully) exogenous supply shifter: bad weather (specifically wind and waves) makes it more difficult to fish, which moves the supply curve inward. (Not needed for this exercise, but if you're curious, see Graddy's fish papers on her website,[1] like page 210 "How the Market Worked at Fulton Street" of her 2006 *JEP* paper.)

a. Load the data with (remove line break)

```
use https://raw.githubusercontent.com/kaplandm/stata/main/data/
   fishdata.dta , clear
```

b. Rename variables to be more intuitive[2]:

```
rename price lnp
rename qty lnq
```

c. Run `reg lnq lnp` and explain why this estimator of the demand curve is not consistent.

d. Explain what it would mean for the weather to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

[1] https://www.kathryngraddy.org/research#pubfish

[2] Unfortunately, there are no variable labels, so there is no way to know these are in logs unless you look back at the original paper.

e. Compute the IV estimator and corresponding tests:
`ivreg2 lnq (lnp=stormy mixed) , robust`

   i. Describe the IV estimand in this example.
   ii. Discuss the economic significance of the estimate.
   iii. Explain what the confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
   iv. Explain why you are or are not worried about weak instruments in this case; refer to specific Stata output.
   v. Explain what the J-test results suggest about the model.

f. Run `ivreg2 lnq (lnp=windspd) , robust` and briefly compare with the previous IV results (slope estimate, weak IV test, J-test).

g. Run

```
ivreg2 lnq (lnp=stormy mixed) , robust gmm2s center
gmm (lnq - lnp*{lnp} - {_cons}) , instruments(stormy mixed) nolog
    vce(robust) twostep
```

   and briefly compare with the slope estimate/CI from part (e).

h. What do you think about a model with a constant slope in this case? That is, a model where the shock/error shifts the demand curve up and down but does not change its slope?

Exercise II.3. The data are originally from Card (1995), with individual-level observations of (log) wages, years of education, and other variables. Note the dataset lacks variable labels, but they can be found online.[3] This is the same dataset as previously in Exercise I.4.

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse card , clear`

c. Create a dummy to compare high-school (12 years education) and college (16 years education):

```
gen d_coll = .
replace d_coll=0 if educ==12
replace d_coll=1 if educ==16
```

d. Regress log wage on years of education `reg lwage educ , vce(robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction.

e. Explain what it would mean for `nearc4` to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

[3] http://fmwww.bc.edu/ec-p/data/wooldridge/card.des

f. Run `ivreg2 lwage (educ = nearc4 nearc2 ) , robust`

  i. Describe the IV estimand in this example.
  ii. Discuss economic significance of the estimated slope ("returns to education").
  iii. Explain what the confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
  iv. Explain why you are or are not worried about weak instruments in this case; refer to specific Stata output.
  v. Explain what the J-test results suggest about the model.

g. Run `ivreg2 lwage (educ = nearc4 ) , robust` and briefly compare with the previous IV results (slope estimate, weak IV test, J-test).

h. Run
`ivreg2 lwage (educ = nearc4) , gmm2s center robust`
`ivreg2 lwage (educ = nearc4 nearc2) , gmm2s center robust`
and comment on differences among these and previous estimates of the return to education.

i. Run `gmm (lwage - educ*{educ} - {_cons}) , instruments(nearc4 nearc2) nolog vce(robust) twostep` and compare with the corresponding estimate/CI from part (h).

j. Run `ivreg2 lwage (d_coll = nearc4 nearc2 ) , robust`

  i. Describe the IV estimand in this example.
  ii. Discuss economic significance of the estimated coefficient on `d_coll`.
  iii. Explain what the confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
  iv. Explain why you are or are not worried about weak instruments in this case; refer to specific Stata output.
  v. Explain what the J-test results suggest about the model.

Exercise II.4. This is another "returns to education" example but with parents' education as the instrument. Note the dataset lacks variable labels, but they can be found online.[4]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse mroz , clear`

c. Regress log wage on years of education `reg lwage educ , vce(robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction.

[4] http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.des

d. Explain what it would mean for `motheduc` to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

e. Run `ivreg2 lwage (educ = motheduc fatheduc) , robust`

   i. Describe the IV estimand in this example.
   ii. Discuss economic significance of the estimated slope ("returns to education").
   iii. Explain what the confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
   iv. Explain why you are or are not worried about weak instruments in this case; refer to specific Stata output.
   v. Explain what the J-test results suggest about the model.

f. Run `ivreg2 lwage (educ = motheduc fatheduc) exper expersq , gmm2s center robust` and briefly compare the estimate and CI for the coefficient on education with the previous estimates/CIs above.

g. Run `gmm (lwage - educ*{educ} - exper*{exper} - expersq*{expersq} - {_cons}) , instruments(motheduc fatheduc exper expersq) nolog vce( robust) twostep` and compare with the estimate/CI from part (f).

Exercise II.5. The following IV analysis uses cigarette prices to instrument for how much a mother smoked while pregnant, in hopes of estimating the causal effect of cigarette smoking on birthweight (which when too low is associated with other negative health outcomes for infants). Note the dataset lacks variable labels, but they can be found online.[5]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse bwght , clear`

c. Run `reg lbwght cigs male parity lfaminc , vce(robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous.

d. Explain what it would mean for `cigprice` to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

e. Run `ivreg2 lbwght (cigs=cigprice) male parity lfaminc , robust`

   i. Describe the IV estimand in this example.
   ii. Discuss the economic significance of the estimated slope on `cigs`.
   iii. Explain what the confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
   iv. Explain why you are or are not worried about weak instruments in this case; refer to specific Stata output.

[5] http://fmwww.bc.edu/ec-p/data/wooldridge/bwght.des

v. Explain what the J-test results suggest about the model.

f. Run `ivreg2 lbwght (cigs=cigprice) male parity lfaminc , robust gmm2s center` and briefly compare with your previous estimate and CI.

g. Run `gmm (lbwght - cigs*{cigs} - male*{male} - parity*{parity} - lfaminc*{lfaminc} - {_cons}) , instruments(cigprice male parity lfaminc) nolog vce(robust) twostep` and briefly compare with your previous estimates/CIs.

Exercise II.6. The following example uses data from Blackburn and Neumark (1992), specifically a cross-section of men in the year 1980, originally from the National Longitudinal Survey (NLS). The analysis uses birth order (1 means first-born in family / oldest child in family; 2 means second-born / second-oldest child in family; etc.) to instrument for how much education someone gets, in hopes of estimating the causal effect of education on (log) wage. Note the dataset lacks variable labels, but they can be found online.[6]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse wage2 , clear`

c. Run `reg lwage educ exper exp2 married , vce(robust)` and explain one potential source of omitted variable bias (for the coefficient on education) along with the direction of bias; be precise and rigorous in your argument for the direction.

d. Explain what it would mean for `brthord` to be a helpful "proxy" variable, and a potential (real-world) reason it may not be.

e. Explain what it would mean for `brthord` to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

f. Run `ivreg2 lwage (educ=brthord) c.exper##c.exper married , robust`

   i. Describe the IV estimand in this example.
   ii. Discuss the economic significance of the estimated slope on `educ`.
   iii. Explain what the confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
   iv. Explain why you are or are not worried about weak instruments in this case; refer to specific Stata output.
   v. Explain what the J-test results suggest about the model.

g. How many of your previous answers would change if we used two-step GMM estimation (instead of IV/2SLS regression)? Explain. (Feel free to re-run the `ivreg2` command with additional options `gmm2s center` to check.)

[6] http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.des

h. Run `gen expersq = exper^2` and then `gmm (lwage - educ*{educ} - exper*{exper} - expersq*{expersq} - married*{married} - {_cons}) , instruments(brthord exper expersq married) nolog vce(robust) twostep` and briefly compare with your previous estimates/CIs.

Part III

Panel Data

Chapter 9

Difference-in-Differences

Unit learning objectives for this chapter

9.1. XXX [TLO 1]

This chapter XXX

Optional resources for this chapter

- XXX
- XXX
- XXX

9.1 Introduction

XXX

Appendix to Chapter 9

9.A Technical Details: XXX

This section shows the technical details for deriving XXX

Exercises

Exercise III.1. You will analyze data on driving laws and fatal accident rates, originally from Freeman (2007). In particular, you'll compare weekend driving fatality (death) rates for states that adopted a 0.08 blood alcohol content (BAC) law and states that didn't, comparing rates before and after the law adoption.

a. Load the data with (remove the line break)

   ```
   use https://raw.githubusercontent.com/kaplandm/stata/main/data/
       driving.dta , clear
   ```

   and read the variable labels (including units of measure): `describe`

b. Keep only years 1980 and 1990: `keep if year==1980 | year==1990`

c. Create an "after" period dummy variable: `gen after = (year==1990)`

d. Create variable `bac` equal to `1` if there's any BAC law that year:
   `gen bac = (bac08 + bac10 >= 1)`

e. Drop states that already had a BAC law in the "before" period (1980), leaving only states that never had the law or adopted it between 1980 and 1990:

   ```
   generate dropflag = ((!after) & bac)
   bysort state : egen dropst = max(dropflag)
   drop if dropst
   ```

f. Create a treatment dummy equal to 1 for states that adopted a BAC law by 1990:
   `bysort state : egen treat = max(bac)`

g. Run a difference-in-differences regression with the intercept, "after" dummy, treatment dummy, and interaction term. Below, the `##` automatically generates the desired interaction term: `reg wkndfatrte treat##after , vce(robust)`

h. To see how the OLS coefficient estimates relate to the conditional means (CMF estimates), compute the sample mean weekend driving fatality rate within each of the four groups defined by the time period and "treatment" status:
   `tabulate treat after , summarize(wkndfatrte) means missing`

i. Display the CMF-based replication of the OLS estimates:

```
collapse (mean) wkndfatrte , by(treat after)
display wkndfatrte[1]
display wkndfatrte[3]-wkndfatrte[1]
display wkndfatrte[2]-wkndfatrte[1]
display (wkndfatrte[4]-wkndfatrte[3])-(wkndfatrte[2]-wkndfatrte[1])
```

j. Repeat part (g) but with a different outcome variable to replace `wkndfatrte`, like the weekend fatalities per 100 million miles driven (instead of population), or the total fatality rate (not just weekends), etc.

k. Repeat parts (d)–(g) but replacing your `bac` treatment variable created in part (d) with a treatment dummy equal to 1 if `perse` (a different driving law) equals 1 (and equal to 0 otherwise).

Exercise III.2. The dataset here has an observation for each state (plus DC) in the U.S. ($i = 1, \ldots, 51$) in years 1987, 1990, and 1993 ($t = 1, 2, 3$). The dependent variable `mrdrte` is the number of murders per 10,000 people (in state i during year t). The `d90` and `d93` are time dummies to include year effects. The two regressors are the unemployment rate (in state i, year t) and the number of executions in state i in years $t-2$, $t-1$, and t combined. (Note: this is not intended to be a sophisticated, definitive analysis upon which you should base your beliefs.)

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`
b. Load the data: `bcuse murder , clear`
c. Run `reg mrdrte d90 d93 exec unem , vce(cluster state)`
d. Run `xtset id year`
e. Run `xtreg mrdrte d90 d93 exec unem, fe cluster(id)`
f. Report the pooled OLS and FE estimated coefficients on `exec`, and explain (both mathematically and in real-world terms) what this suggests about the relationship between `exec` and the unobserved state effects.
g. Discuss the economic significance of the FE estimated coefficient.
h. Explain what the corresponding confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
i. Think of one additional (unobserved) time-varying variable that might also be correlated with `exec`. Explain which sign (positive or negative) you think the correlation might have, and in which direction this would bias the FE estimator.

Exercise III.3. The following dataset is not a panel but a repeated cross-section that includes years 1978 and 1981 ($t = 1, 2$), between which a new garbage incinerator was built in a particular neighborhood. Interest is in the causal effect on house prices; variable `lrprice` has log real house prices. Note `y81` is a dummy for year 1981, and `nearinc` is a dummy for being "near" the incinerator's location (even if it's 1978 and the incinerator itself does not yet exist).

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse kielmc , clear`

c. Run `reg lrprice nearinc if y81 , vce(robust)` and say what that code estimates as well as a specific real-world reason you think this is a biased estimator of the causal effect of being near the incinerator on housing price.

d. Run `reg lrprice y81 if nearinc , vce(robust)` and say what that code estimates as well as a specific real-world reason you think the estimator is biased.

e. Run `reg lrprice nearinc##y81 , vce(robust)`

   i. Report the number that is the difference-in-differences estimator of the effect of interest, as well as the units of measure.

   ii. What is the population estimand of this diff-in-diff estimator? Provide both math and real-world description (including definitions of the potential outcomes).

   iii. Discuss the economics significance of the estimate.

   iv. Explain what the confidence interval tells us about our uncertainty about the true population value: be precise and explicit.

   v. Recall that here we only have a repeated cross-section (not panel), and house prices are only observed when a house is sold. Assume conditions are relatively normal, so houses not near the incinerator (`nearinc=0`) are essentially sold at random (somebody gets a job in another state, somebody moves into a retirement home, etc.), so our dataset has a random sample of such house prices, and similarly in 1978 for all houses. Why might the 1981 near-incinerator prices not be a random sample, i.e., why might those houses not just be sold randomly? In which direction might this bias the diff-in-diff estimator? (There are many possible aspects to consider, but if you're having trouble getting started: imagine usually 5% of houses in a neighborhood sell in a typical year; 20% of homeowners are extremely opposed to living near a garbage incinerator while 80% don't care at all; recall basic supply and demand, how price responds to an increase in supply that shifts the supply curve; etc.)

Exercise III.4. The following analyzes county-year level crime data from North Carolina. The variable descriptions can be found online.[1] (Some of the descriptions are still vague; for research you would want to understand the variables much better, but we'll focus on other issues for now.)

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse crime4 , clear`

c. Run `xtset county year`

[1] http://fmwww.bc.edu/ec-p/data/wooldridge/crime4.des

d. Run `reg lcrmrte lpolpc if year==87 , vce(robust)` and explain one specific reason you don't think the slope coefficient can be interpreted as a causal effect; say in which direction you think it is biased, and why.

e. Run `reg lcrmrte lpolpc d8* , vce(robust)` and explain why this does not address your above concern (or if it does, come up with a different reason you don't think this estimates a causal effect).

f. Run `xtreg lcrmrte lpolpc d8* , fe cluster(county)`

   i. Explain what type of omitted variable (bias) the county-level fixed effects capture.
   ii. Discuss the economic significance of the FE estimated coefficient on `lpolpc`.
   iii. Explain what the corresponding confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
   iv. Explain why this FE model still does not identify a causal effect in this example, including the direction of bias. (Feel free to try `reg lpolpc lcrmrte d8*` while you're thinking.)

Exercise III.5. The following analyzes data on manufacturing scrap rates for firms that did or did not receive grant money to improve. The variable descriptions can be found online.[2]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse jtrain , clear`

c. Run `xtset fcode year`

d. Run `reg grant L.lscrap` and briefly say what this suggests about which firms receive a grant. (Note: for real research, you would want to read about the grant program itself, not just run a simple regression.)

e. Run `reg lscrap L.lscrap` and briefly say what this suggests about firms' scrap rates over time.

f. Run `reg lscrap grant grant_1 if year==1989 , vce(robust)` and explain one specific reason you don't think the slope coefficient can be interpreted as a causal effect; say in which direction you think it is biased, and why. (Hint: think about your previous two results.)

g. Run `xtreg lscrap grant grant_1 d88 d89 , fe cluster(fcode)`

   i. Explain what type of omitted variable (bias) the firm-level fixed effects capture.
   ii. Discuss the economic significance of the FE estimated coefficients on `grant` and `grant_1`.

[2] http://fmwww.bc.edu/ec-p/data/wooldridge/jtrain.des

iii. Explain what the corresponding confidence intervals tell us about our uncertainty about the true population values; be precise and explicit.
iv. Explain what would need to be true for strict exogeneity to be satisfied here.

h. Run `lincom grant + grant_1` to get the estimate and confidence interval for the sum of these coefficients; how do you interpret this sum economically?
i. Run `reg D(lscrap grant grant_1 d89) , vce(cluster fcode)` to compute the FD estimator and briefly compare with the FE results.

Exercise III.6. The following analyzes crime data from Norway. The "clear-up percentage" is how many reported crimes were resolved by charging an individual with the crime (most commonly), which may be a deterrent to future crime. The variable descriptions can be seen in the variable labels.

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`
b. Load the data: `bcuse norway , clear`
c. Run `xtset district year , delta(6)` noting that the `delta(6)` tells it to treat year 1972 as $t = 1$ and 1978 as $t = 2$.
d. Run `reg lcrime clrprc1 clrprc2 if year==78 , vce(robust)` and explain one specific reason you don't think the slope coefficient can be interpreted as a causal effect; say in which direction you think it is biased, and why.
e. Run `xtreg lcrime clrprc1 clrprc2 d78 , fe cluster(district)`

   i. Explain what type of omitted variable (bias) the district-level fixed effects capture.
   ii. Discuss the economic significance of the FE estimated coefficients on `clrprc1 and clrprc2`.
   iii. Explain what the corresponding confidence intervals tell us about our uncertainty about the true population values; be precise and explicit.
   iv. Explain one possible reason that strict exogeneity might be violated here.

f. Run `reg D(lcrime clrprc1 clrprc2) , vce(cluster district)` to compute the FD estimator and briefly compare with the FE results.

Exercise III.7. The following examines the relationship between low infant birthweight (a bad health outcome) and participation in a welfare program (that hopes to help pregnant women through nutrition programs and prenatal care). The specific program is the Aid to Families with Dependent Children (AFDC). The panel data is aggregated at the state-year level. Other control variables try to proxy for general quality of health care and income level in the state. The variable descriptions can be found online.[3]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

[3] `http://fmwww.bc.edu/ec-p/data/wooldridge/lowbrth.des`

b. Load the data: `bcuse jtrain , clear`

c. Run `encode state , gen(state_id)` to get a numeric identifier for the states (because `xtset` does not allow strings).

d. Run `xtset state_id year , delta(3)` noting that the `delta(3)` tells it to treat year 1987 as $t = 1$ and 1990 as $t = 2$.

e. Run `reg lowbrth afdcprc if year==1990 , vce(robust)` and explain one specific reason you don't think the slope coefficient can be interpreted as a causal effect; say in which direction you think it is biased, and why.

f. Run `xtreg lowbrth afdcprc d90 , fe cluster(state_id)`

   i. Explain what type of omitted variable (bias) the state-level fixed effects capture.
   ii. Discuss the economic significance of the FE estimated coefficient on `afdcprc`.
   iii. Explain what the corresponding confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
   iv. Explain one possible reason that strict exogeneity might be violated here.

g. Run `reg D(lowbrth afdcprc) , vce(cluster state_id)` to compute the FD estimator and briefly compare with the FE results.

h. Run `xtreg lowbrth afdcprc d90 lphypc lbedspc lpcinc lpopul , fe cluster(state_id)` and briefly compare with the earlier FE results that did not include control variables.

Part IV

Probit

Introduction

This part concerns XXX

Chapter 10

Binary Response Models

Unit learning objectives for this chapter

10.1. XXX [TLO 1]

Optional resources for this chapter

- XXX
- XXX
- XXX
- XXX

10.1 XXX

Exercises

Exercise IV.1. Consider the binary variable (`inlf` below) of whether or not a married woman is in the labor force, and its relationship with other socioeconomic variables. Note the dataset lacks variable labels, but they can be found online.[1]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`
b. Load the data: `bcuse mroz , clear`
c. Run `reg inlf educ exper kidslt6 kidsge6 nwifeinc , vce(robust)`
    i. Describe how to interpret the population model you are estimating.
    ii. Interpret the estimated coefficient on `educ`, and comment on its economic significance.
    iii. Explain what the confidence interval tells us about our uncertainty about the true population value; be precise and explicit.
    iv. Explain one reason (specific to this economic example) that you doubt the true conditional mean function is linear-in-variables.
d. Run `probit inlf educ exper kidslt6 kidsge6 nwifeinc , vce(robust)`
e. Run `margins, dydx(educ exper) atmeans` and `margins, dydx(educ exper)` and explain the difference between the two commands; then compare the results with the OLS estimated slopes.
f. Run `logit inlf educ exper kidslt6 kidsge6 nwifeinc` followed by `margins, dydx(educ exper)` and briefly comment on the economic significance of the difference with the probit-estimated average partial "effects."
g. Consider the following very stylized hypothetical predication application. Imagine you work for a company that offers services for married women in the labor force, and your job is to write code to decide whether or not to buy an online ad for each user that visits another website (that allows you to buy ads for a fixed price). The other website collects all the regressors (predictors) used above, but cannot observe

[1] http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.des

`inlf`, so you need to guess (predict). Given your estimated model from above, you can compute the predicted (conditional) probability of being in the labor force for user i, denoted $\hat{p}_i$. Each ad costs \$0.001; if the user is indeed in the labor force, then expected revenue is \$0.003 (because most people don't click through, etc.), otherwise expected revenue is zero. Assuming your goal is to maximize expected profit, which is a better prediction of being in the labor force (y_i), $\hat{y}_i = \mathbb{1}\{\hat{p}_i > 0.25\}$ or $\hat{y}_i = \mathbb{1}\{\hat{p}_i > 0.75\}$? (That is, you generate binary $\hat{y}_i$, then run the ad if $\hat{y}_i = 1$ but not if $\hat{y}_i = 0$.) Try to find an even better prediction rule for $\hat{y}_i$ as a function of $\hat{p}_i$, and explain why your prediction generates higher expected profit than the two above.

Exercise IV.2. The following models whether an individual is arrested or not in a particular year, given their past criminal justice involvement, demographics, and current employment and income. Variable descriptions are provided in the variable labels in the dataset, originally studied by Grogger (1991). Section II of the original paper provides more details about the data, like covering men in California who were arrested at least once since 1972 and who were born in either 1960 or 1962.

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`
b. Load the data: `bcuse grogger , clear`
c. Create the dependent variable: `gen d_arr86 = (narr86>0)`
d. Run `reg d_arr86 pcnv avgsen tottime black hispan , vce(robust)`
    i. Describe how to interpret the population model you are estimating.
    ii. Interpret the estimated coefficients on `pcnv` and `avgsen`, and comment on their economic significance.
    iii. Explain what the confidence intervals tell us about our uncertainty about the true population values; be precise and explicit.
    iv. Explain one reason (specific to this economic example) that you doubt the true conditional mean function is linear-in-variables.
e. Run `probit d_arr86 pcnv avgsen tottime black hispan , vce(robust)`
f. Run `margins, dydx(pcnv avgsen) atmeans` and `margins, dydx(pcnv avgsen)` and explain the difference between the two commands; then compare the results with the OLS estimated slopes.
g. Run `logit d_arr86 pcnv avgsen tottime black hispan` followed by `margins , dydx(pcnv avgsen)` and briefly comment on the economic significance of the difference with the probit-estimated average partial "effects" of `pcnv` and `avgsen`.
h. Now consider trying to predict whether or not an individual will be arrested over the next 12 months for the purpose of targeting an intervention that includes 1-on-1 mentoring, job training, and subsidized housing, and imagine you only care about reducing arrests (not any other outcome). There is no budget constraint, but

the opportunity cost of spending \$1 on this program is not spending that \$1 on a different program to help reduce arrests. After running your `probit` command from above, pretend we then loaded a new dataset that includes only the predictor variables but not `d_arr86`, and then generate the predicted arrest probabilities with `predict phat` along with two possible binary predictions of being arrested using two different probability thresholds:

```
gen target25 = (phat>0.25)
gen target48 = (phat>0.48)
```

Finally, because actually we do know the true `d_arr86` values, compare the true and predicted values:

```
tab d_arr86 target25
tab d_arr86 target48
```

i. For the 25% threshold: how many "false negatives" (`target25=0` but they are arrested) and "false positives" (`target25=1` but they are not arrested) are there? How many are there for the 48% threshold?
ii. Qualitatively, what is the cost of a false negative? What's the cost of a false positive?
iii. Adding whatever additional details you need (about costs, benefits, etc.) for the following to be true: why might the higher threshold be preferred here?
iv. Would a 50% threshold be even better? 60%? Explain why/not.

Exercise IV.3. Consider the relationship between whether or not somebody reports being in good health (`gdhlth`) and other variables. This dataset is from 1975. Note the dataset lacks variable labels, but they can be found online.[2]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`
b. Load the data: `bcuse sleep75 , clear`
c. Run `reg gdhlth c.age##c.age male##yngkid sleep totwrk educ , vce(robust)`
    i. Describe how to interpret the population model you are estimating.
    ii. Interpret the estimated coefficients on `age` and its square, and comment on their economic significance.
    iii. Explain what the confidence intervals tell us about our uncertainty about the true population values; be precise and explicit.
    iv. Explain one reason (specific to this economic example) that you doubt the true conditional mean function has this exact functional form.
d. Run `margins , dydx(age) at(age=(30(15)60)) vsquish`

[2] http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.des

e. Run `probit gdhlth c.age##c.age male##yngkid sleep totwrk educ , vce(robust)` and then `margins , dydx(age) at(age=(30(15)60)) vsquish` and compare with the OLS results.

f. Repeat part (e) but with `logit` instead of `probit` and briefly compare to the `probit` results.

g. Now imagine you work for a health insurance company and want to predict if an individual is in good health; if not, the insurance company will call them with a reminder to visit the doctor. After running your `probit` command from above, pretend we then loaded a new dataset that includes only the predictor variables but not `gdhlth`, and then generate the predicted arrest probabilities with `predict phat` along with two possible binary predictions of being arrested using two different probability thresholds:
```
gen target50 = (phat>0.50)
gen target80 = (phat>0.80)
```
Finally, because actually we do know the true `gdhlth` values, compare the true and predicted values:
```
tab gdhlth target50
tab gdhlth target80
```

   i. For the 50% threshold: how many extraneous phone calls would be made (`target50=0` but `gdhlth=1`)? How many individuals not in good health fail to get called (`target50=1` but `gdhlth=0`)? How many of each for the 80% threshold?

   ii. Qualitatively, what is the cost of calling somebody who's actually in good health? What's the cost of failing to call somebody in bad health?

   iii. Adding whatever additional details you need (about costs, benefits, etc.) for the following to be true: why might the higher 80% threshold be preferred here?

Bibliography

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1):265–296. URL https://doi.org/10.3982/ECTA12675. [19]

Baum, Christopher F. 2012. "BCUSE: Stata module to access instructional datasets on Boston College server." Statistical Software Components S457508, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s457508.html. Revised 12 Nov 2021. [9]

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2002. "IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation." Statistical Software Components S425401, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s425401.html. Revised 10 May 2022. [9]

Blackburn, McKinley and David Neumark. 1992. "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials." *Quarterly Journal of Economics* 107 (4):1421–1436. URL https://www.jstor.org/stable/2118394. [82]

Card, David. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, edited by Louis N. Christophides, E. Kenneth Grant, and Robert Swidinsky. University of Toronto Press, 201–222. [66, 79]

Deming, W. Edwards and Frederick F. Stephan. 1941. "On the Interpretation of Censuses as Samples." *Journal of the American Statistical Association* 36 (213):45–49. URL https://www.jstor.org/stable/2278811. [19, 20]

Freeman, Donald G. 2007. "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws." *Contemporary Economic Policy* 25 (3):293–308. URL https://doi.org/10.1111/j.1465-7287.2007.00039.x. [91]

Graddy, Kathryn. 1995. "Testing for imperfect competition at the Fulton fish market." *RAND Journal of Economics* 26 (1):75–92. URL https://www.jstor.org/stable/2556036. [78]

Grogger, Jeffrey. 1991. "Certainty vs. Severity of Punishment." *Economic Inquiry* 29 (2):297–309. URL https://doi.org/10.1111/j.1465-7295.1991.tb01272.x. [104]

Hansen, Bruce E. 2020a. "Econometrics." URL https://www.ssc.wisc.edu/~bhansen/

econometrics. Textbook draft. [xi, 3, 15]

———. 2020b. "Introduction to Econometrics." URL https://www.ssc.wisc.edu/~bhansen/probability. Textbook draft. [xi, 3]

Imbens, Guido and Jeffrey M. Wooldridge. 2007. "What's New in Econometrics: Estimation of Average Treatment Effects Under Unconfoundedness." NBER summer lecture notes, available at https://www.nber.org/WNE/lect_1_match_fig.pdf. [48]

Kaplan, David M. 2021. "Distributional and Nonparametric Econometrics." URL https://kaplandm.github.io/teach.html. Textbook draft. [xi, 13, 22, 29, 33, 35, 43, 49, 57]

———. 2022. *Introductory Econometrics: Description, Prediction, and Causality.* Columbia, MO: Mizzou Publishing, 3rd ed. URL https://www.themizzoustore.com/p-236916-introductory-econometrics-description-prediction-and-causality.aspx. [xi, 13, 17, 18, 21, 22, 25, 29, 30, 32, 37, 40, 45, 46, 47, 48]

Kaplan, David M. and Longhao Zhuo. 2021. "Frequentist properties of Bayesian inequality tests." *Journal of Econometrics* 221 (1):312–336. URL https://doi.org/10.1016/j.jeconom.2020.05.015. [25]

Kleibergen, Frank, Mark E. Schaffer, and Frank Windmeijer. 2007. "RANKTEST: Stata module to test the rank of a matrix." Statistical Software Components S456865, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s456865.html. Revised 29 Sep 2020. [9]

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4):604–620. URL https://www.jstor.org/stable/1806062. [65]

Lewbel, Arthur. 2019. "The Identification Zoo: Meanings of Identification in Econometrics." *Journal of Economic Literature* 57 (4):835–903. URL https://doi.org/10.1257/jel.20181361. [31]

Lucas, Robert E., Jr. 1976. "Econometric policy evaluation: A critique." In *Carnegie–Rochester Conference Series on Public Policy*, vol. 1. North-Holland, 19–46. [30]

StataCorp. 2017. "Stata Statistical Software: Release 15." College Station, TX: StataCorp LP. [9]

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT Press, 2nd ed. URL https://www.worldcat.org/oclc/831625495. [32, 45, 61]

———. 2020. *Introductory Econometrics: A Modern Approach.* Cengage, 7th ed. [65]

Index

ACE, *see* average causal effect
after sampling, 21
all-causes model, 52, 58
analogy principle, 29
ASE, *see* average structural effect
ASF, *see* average structural function
asymptotic bias, 43
ATE, *see* average treatment effect
average causal effect, 47
average structural effect, 53
average structural function, 52
average treatment effect, 47

Bayesian, 25
before sampling, 21
best linear approximation, 33
best linear predictor, 32
bias, 40
    attenuation, 40
    downward, 40
    negative, 40
    positive, 40
    toward zero, 40
    upward, 40
BLA, *see* best linear approximation
BLP, *see* best linear predictor

CATE, *see* conditional average treatment effect
causal inference, 30
CEF, *see* conditional expectation function
CI, *see* confidence interval
classical, 25
CMF, *see* conditional mean function
conditional average treatment effect, 55
conditional expectation function, 33
conditional mean function, 33
confidence interval, 37
confidence level, 38
consistent, 43
contrapositive, 15, 16
converse, 15
counterfactual, 30
covariates, 32
coverage probability, 38
    nominal, 38
CP, *see* coverage probability
credible interval, 26
credible set, 26

data-generating process, 19
DGP, *see* data-generating process

economic significance, 36
efficient, 43
empirical distribution, 18
endogenous, 51
error form, 34
exogenous, 51

frequentist, 25

GE, *see* general equilibrium

general equilibrium, 29
general equilibrium effects, 50

heterogeneity, 46

identically distributed, 23
identification, 28
- nonparametric, 56
- partial, 28
- point, 28
- set, 28

identified, 28
identifying assumptions, 28
if, 14
if and only if, 14
ignorability, 48
iid, *see* independent and identically distributed
implied by, 14
implies, 14
independence, 48
- conditional mean, 56
- mean, 49

independent and identically distributed, 23
independent variables, 32
inverse, 15

linear projection, 32
linear projection coefficients, 32
LP, *see* linear projection
LPCs, *see* linear projection coefficients

mean squared error, 42
MSE, *see* mean squared error
multiple comparisons problem, 39
multiple testing problem, 39

necessary, 14
Neyman–Rubin causal model, 45
no interference, 48
non-interference, 48
nonparametric regression, 35
nonseparable, 52

OLS, *see* ordinary least squares
only if, 14
ordinary least squares, 32

partial equilibrium, 29
PE, *see* partial equilibrium
plug-in principle, 29
population
- finite, 19
- infinite, 19
- super-, 19

posterior, 25
potential outcome
- treated, 46
- untreated, 46

predictors, 32
prior, 25

quadratic loss, 33

random
- -ized, 30
- draw, 21
- sample, 21, 23
- variable, 21

random coefficients, 51
realization, 21
realized value, 21
reduced-form, 30
regression discontinuity, 58
regressors, 32
repeated sampling, 26
right-hand-side variables, 32

sample
- analog, 34
- distribution, *see* empirical distribution
- size, 22

sampling
- independent, 23

sampling distribution, 40
significance
- economic, *see* economic significance

statistical, *see* statistical significance
spillover effects, 50
statistically significant, 37
statistics, 2
strong ignorability, 48
stronger, 14
structural approach, 30
sufficient, 14

treatment, 45
treatment effect, 45, 46

unbiased, 40
unconfoundedness, 48
units, 22

weaker, 14