Theory and Methods in Causal Inference

Lecture Notes

Jae Kwang Kim

Department of Statistics

Iowa State University

STAT 5900B

Last updated: April 29, 2026

Contents

Preface

These lecture notes accompany the graduate course **STAT 5900B: Theory and Methods in Causal Inference** at the Department of Statistics, Iowa State University.

The course targets first-year PhD students in statistics. It assumes familiarity with probability theory, mathematical statistics, and linear models at the level of a graduate sequence, but no prior exposure to causal inference.

Download full PDF

Organization

The notes are organized in three parts, with a fourth part on policy learning deferred to a later release.

Part I — Foundations (Chapters 1–3) builds the graphical and conceptual foundation of causal inference. Chapter 1 introduces the three languages of causality — structural equation models (SEMs), directed acyclic graphs (DAGs), and potential outcomes — and explains how they relate. Chapter 2 develops DAGs and d-separation as tools for reading off conditional independence from causal structure. Chapter 3 introduces the do-calculus and the main identification criteria (back-door, front-door, do-calculus rules) that translate causal queries into statistical ones.

Part II — Identification (Chapters 4–9) bridges the graphical framework and observed data. Chapter 4 connects the potential outcomes framework to the do-calculus and establishes the consistency, exchangeability, and positivity assumptions. Chapter 5 covers randomization and the back-door adjustment formula. Chapter 6 develops propensity score methods — weighting, matching, and the balancing property. Chapter 7 develops instrumental variables: the Wald estimand, the local average treatment effect (LATE) for compliers, and the three IV assumptions. Chapter 8 covers mediation analysis and the front-door criterion, distinguishing the controlled direct effect from the natural direct and indirect effects. Chapter 9 addresses sensitivity analysis and partial identification: the three canonical sensitivity models (Rosenbaum Γ, E-value, marginal sensitivity model), Manski's no-assumption bounds, and the connection between sensitivity analysis and modern estimation methods.

Part III — Estimation (Chapters 10–13) covers semiparametric efficiency theory and its modern applications. Chapter 10 introduces the estimation framework — estimating equations, asymptotic linearity, influence functions, Z-estimation with nuisance parameters, and the semiparametric efficiency bound — that underlies the rest of Part III. Chapter 11 develops the augmented inverse probability weighted (AIPW) estimator from three complementary perspectives — bias correction, optimal augmentation, and the efficient influence function — and establishes double robustness and semiparametric efficiency. Chapter 12 addresses flexible nuisance estimation with orthogonal scores and cross-fitting, culminating in the double machine learning (DML) estimator and its rate conditions. Chapter 13 covers estimation under instrumental variables: the Wald estimator, two-stage least squares, GMM and overidentification tests, weak instruments, generalized empirical likelihood, and the control function approach.

Part IV — Policy Learning and Sequential Decision Making (Chapters 14–16, deferred) will cover policy evaluation, optimal policy learning, and dynamic treatment regimes.

Appendices

Appendix A provides graphical intuition for conditional independence and d-separation. **Appendix B** introduces Single World Intervention Graphs (SWIGs). **Appendix C** develops the formal foundations of pathwise differentiability and efficient influence functions for readers who want the semiparametric geometry underlying Chapter 10.

Notation

Throughout these notes:

- $\text{do}(T = t)$ denotes the do-operator (intervention).
- $Y(t)$ denotes the potential outcome under $T = t$.
- $\mathbb{E}[\,\cdot\,]$ denotes expectation; $\perp\!\!\!\perp$ denotes conditional independence.
- $\mathcal{G}$ denotes a DAG; $\mathcal{G}_{\overline{T}}$ denotes the mutilated graph after intervening on T.
- $\pi(x) = P(T = 1 \mid X = x)$ denotes the propensity score.
- $\mu_t(x) = \mathbb{E}(Y \mid T = t, X = x)$ denotes the outcome regression.

Acknowledgements

I would like to thank professors Chan Park, Shu Yang, and Dylan Small for their constructive comments on the previous version of this lecture note.

Part I

Foundations

Chapter 1

Introduction

Learning Objectives

By the end of this chapter, students should be able to:

1. Articulate the distinction between an interventional distribution $P(y \mid \text{do}(T{=}t))$ and an observational conditional distribution $P(y \mid T{=}t)$, and explain why they differ whenever T is endogenous.
2. Describe the same causal question in all three languages — SEM, DAG graph surgery, and potential outcomes — and begin to translate between them.
3. Derive the two densities explicitly in the Gaussian linear confounded model, and identify the endogeneity bias $\alpha\gamma\sigma_U^2/\sigma_T^2$ as the gap between them.
4. State the two-step paradigm (identification then estimation) and locate each step in the causal inference pipeline.

This chapter is an orientation, not a complete treatment. It introduces four ideas — the interventional/observational distinction, the causal trinity, the Gaussian confounded model as a running example, and the identification-versus-estimation paradigm — that will be developed rigorously in the chapters that follow.

1.1 Motivation: Why Causal Inference Is Hard

1.1.1 The Central Question

Causal inference is concerned with a deceptively simple question: what would happen to Y if we were to *set* $T = t$? This is fundamentally different from the question that standard statistical procedures answer: what is the distribution of Y among units *observed* to have $T = t$?

Definition: Interventional vs. Observational Distribution

- The *interventional distribution* $P(y \mid \text{do}(T{=}t))$ is the distribution of Y in a hypothetical world where T has been *externally set* to t, regardless of the usual data-generating mechanism for T.
- The *observational conditional distribution* $P(y \mid T{=}t)$ is the distribution of Y among the subpopulation of units for whom $T = t$ was *observed* to hold.

These two distributions coincide when T is *exogenous* — that is, when T shares no common causes with Y. Whenever T is endogenous, the two distributions differ, and the difference is *endogeneity bias*.

Example: Drug Trial with Unmeasured Severity

Does a new drug ($T = 1$) reduce blood pressure (Y)? Suppose sicker patients are more likely to take the drug, so unmeasured severity U is a common cause of T and Y. Then:

- *Observational study*: $\mathbb{E}[Y \mid T{=}1] > \mathbb{E}[Y \mid T{=}0]$. Drug users have higher blood pressure on

average. A naïve analyst concludes the drug is harmful.
- *Randomized trial*: $\mathbb{E}[Y \mid \text{do}(T{=}1)] < \mathbb{E}[Y \mid \text{do}(T{=}0)]$. Forcing everyone to take the drug reduces blood pressure. The drug is beneficial.

Same data. Different questions. Opposite answers. The entire discrepancy is caused by the back-door path $T \leftarrow U \rightarrow Y$.

The Fundamental Problem of Causal Inference

We observe $P(y \mid T{=}t)$ directly from data. We want $P(y \mid \text{do}(T{=}t))$. The goal of *identification theory* is to find conditions under which the former can be used to recover the latter.

1.2 The Causal Trinity: Three Languages for One Idea

The same causal question can be expressed in three closely related languages: the *structural equation model* (SEM), the *directed acyclic graph* (DAG) with graph surgery, and the *potential outcomes* framework. Collectively, we call these the **causal trinity**. Each language illuminates a different facet of the same underlying idea, and fluency in all three is essential for modern causal reasoning.

```
\usetikzlibrary{arrows.meta, positioning, calc, shapes.geometric}
\newcommand{\Gcal}{\mathcal{G}}
\definecolor{isublue}{RGB}{30,56,100}
\definecolor{accent}{RGB}{46,117,182}
\definecolor{defbg}{RGB}{238,244,251}
\definecolor{darkgrey}{RGB}{80,80,80}
\definecolor{semcolor}{RGB}{30,56,100}
\definecolor{dagcolor}{RGB}{26,122,58}
\definecolor{pocolor}{RGB}{150,50,150}
\begin{tikzpicture}[
  lang/.style={rectangle,rounded corners=6pt,draw,very thick,text width=3.0cm,align=center,minimum h
  sem/.style={lang,draw=semcolor,fill=semcolor!10,text=semcolor},
  dag/.style={lang,draw=dagcolor,fill=dagcolor!10,text=dagcolor},
  po/.style={lang,draw=pocolor,fill=pocolor!10,text=pocolor},
  center/.style={ellipse,draw=accent,fill=defbg,thick,text width=2.6cm,align=center,font=\small\itsha
  trans/.style={<->,thick,color=darkgrey,shorten >=4pt,shorten <=4pt},
  tlabel/.style={font=\footnotesize\itshape,color=darkgrey,fill=white,inner sep=2pt}
]
\node[po]  (PO)  at (90:3.8cm)   {Potential\\Outcomes\\$Y(t)$};
\node[sem] (SEM) at (210:3.8cm)  {SEM\\$Y=f_Y(T,U,\varepsilon)$};
\node[dag] (DAG) at (330:3.8cm)  {DAG \&\\Graph Surgery\\$\Gcal,\;\Gcal_{\overline{T}}$};
\node[center] (C) at (0,0) {Set $T=t$?};
\draw[-{Stealth[length=5pt]},thick,color=pocolor!70,shorten >=4pt,shorten <=4pt] (PO)--(C);
\draw[-{Stealth[length=5pt]},thick,color=semcolor!70,shorten >=4pt,shorten <=4pt] (SEM)--(C);
\draw[-{Stealth[length=5pt]},thick,color=dagcolor!70,shorten >=4pt,shorten <=4pt] (DAG)--(C);
\draw[trans] (PO)--(SEM) node[tlabel,midway,left=2pt]{\parbox{2.2cm}{\centering consistency\\+ no in
\draw[trans] (SEM)--(DAG) node[tlabel,midway,below=4pt]{\parbox{2.2cm}{\centering equation\\$\leftri
\draw[trans] (DAG)--(PO) node[tlabel,midway,right=2pt]{\parbox{2.2cm}{\centering SUTVA\\+ graph stru
\end{tikzpicture}
```

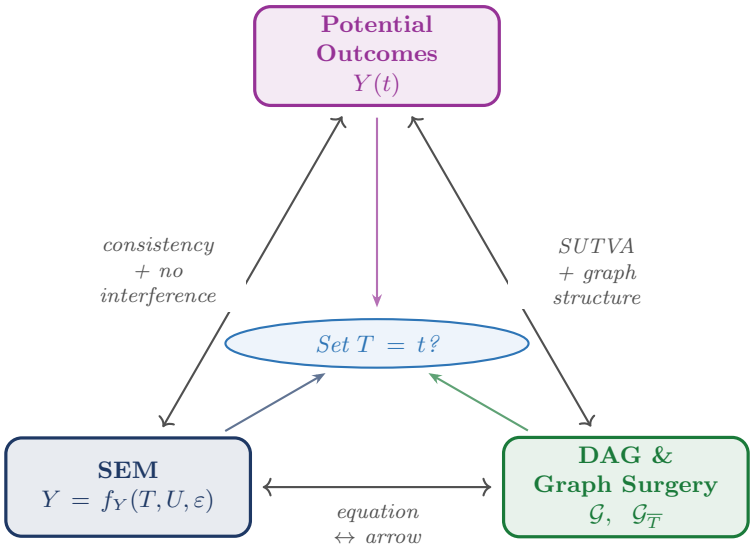


Figure 1.1: The **causal trinity**: three languages that express the same causal question from different vantage points.

1.2.1 Language 1: Structural Equation Models

A SEM represents the causal data-generating mechanism as a system of equations, one per variable, with a joint distribution over exogenous disturbances. A variable is *exogenous* if it is determined outside the system; a variable is *endogenous* if its value is produced by a structural equation within the model.

The running setup for this chapter is a three-variable confounded SEM over (U, T, Y):

$$T = f_T(U, \delta), \qquad Y = f_Y(T, U, \varepsilon), \tag{1.1}$$

where δ and ε are mutually independent exogenous disturbances and U is an unobserved exogenous confounder.

> Remark: Error Independence in the NPSEM-IE
>
> The assumption that the disturbances δ and ε are *mutually independent* is a substantive restriction. Because U is unobserved and enters both f_T and f_Y, error independence does not remove the T–$Y(t)$ dependence induced by U; it restricts only the additional channel through the equation-specific errors (δ, ε). Richardson and Robins (2014) call SEMs with this structure *Nonparametric Structural Equation Models with Independent Errors* (NPSEM-IEs). These notes adopt the NPSEM-IE framework throughout: it ensures the Markov condition holds and integrates cleanly with the graphical identification theory of later chapters.

Observational regime. Because U enters both equations, conditioning on T changes the distribution of U, so $P(y \mid T{=}t) \neq P(y \mid \mathrm{do}(T{=}t))$. An analyst who regresses Y on T without observing U faces a residual correlated with T — even though the structural error ε is independent of T by NPSEM-IE.

Interventional regime. The do-operator replaces the equation for T with a constant t, yielding the *mutilated system*: $U \sim p(u)$, $T = t$ (fixed), $Y = f_Y(t, U, \varepsilon)$. We define the *potential outcome* as $Y(t) := f_Y(t, U, \varepsilon)$; by construction, $Y(t) \sim P(y \mid \mathrm{do}(T{=}t))$.

1.2.2 Language 2: Directed Acyclic Graphs

> Definition: Graph Surgery
>
> In the observational DAG, arrows into T encode the mechanisms that determine T. Under $\mathrm{do}(T{=}t)$, all arrows into T are *deleted*, producing the mutilated graph $\mathcal{G}_{\overline{T}}$. This severs spurious back-door paths while preserving directed causal pathways from T to its descendants.

1.2.3 Language 3: Potential Outcomes

Definition: Potential Outcome [@rubin1974estimating]

In the SEM framework, $Y(t) = f_Y(t, U, \varepsilon)$ is the outcome that *would have been observed* had T been set to t, possibly contrary to fact.

Definition: SUTVA

The *stable unit treatment value assumption* (SUTVA) has two components: (i) *no interference* — unit i's potential outcome depends only on unit i's own treatment; and (ii) *no hidden versions of treatment* — each treatment level corresponds to a single, well-defined intervention. Under SUTVA, the *consistency* relation $Y_i = Y_i(T_i)$ holds. SUTVA is developed further in Chapter 4.

1.2.4 Roadmap: How the Trinity Organizes These Notes

Part I — Foundations: the graphical language (Chapters 1–3). The DAG and the do-operator are the primary language here because they make the interventional/observational distinction *syntactically explicit*: confounding is visible as an open back-door path, and identifying assumptions are visible as graph criteria.

Part II — Identification: designs and research strategies (Chapters 4–9). Part II introduces the potential outcomes framework (Chapter 4) and then studies the specific research designs through which observational and experimental data support identification: randomization and back-door adjustment (Chapter 5), propensity score methods (Chapter 6), instrumental variables (Chapter 7), mediation and front-door identification (Chapter 8), and sensitivity analysis (Chapter 9). Ignorability ($Y(t) \perp\!\!\!\perp T \mid X$) is the potential-outcomes counterpart of the back-door criterion.

Part III — Estimation: semiparametric and machine-learning methods (Chapters 10–13). Once identification has been established, Part III turns to estimation. The semiparametric efficiency framework — estimating equations, influence functions, doubly robust estimators, and double machine learning — is largely language-neutral. Chapter 13 applies these tools to estimation under instrumental variables.

1.2.5 Historical Roots of the Causal Trinity

The three languages did not emerge from a single research program but grew independently in different disciplines over more than a century.

Structural Equation Models: genetics and econometrics. Sewall Wright (1921) introduced *path analysis* to decompose correlations among hereditary traits. The framework was transplanted into economics by Haavelmo (1943), who argued that economic relationships should be modeled as autonomous structural equations robust to policy interventions. The Cowles Commission formalized simultaneous equation systems throughout the 1940s–50s. Lucas's (1976) critique is essentially a statement of the interventional/observational distinction, decades before that phrase existed.

Potential Outcomes: experimental statistics and epidemiology. Neyman (1923) introduced the notation $Y_i(t)$ in the context of randomized agricultural experiments. The crucial extension to *observational* studies was made by Rubin (1974), who formalized the assignment mechanism and stated SUTVA explicitly. Holland's (1986) *JASA* paper brought the framework to mainstream statistics. Robins (1986) extended potential outcomes to longitudinal treatments via g-computation. Imbens and Angrist (1994) connected the framework to instrumental variables and defined the LATE.

DAGs and do-calculus: computer science and philosophy. Pearl developed Bayesian networks through the 1980s. The pivotal step came in Pearl (1995), where the *do-operator* was introduced. The 2000 monograph *Causality* synthesized DAGs, the do-calculus, identification theory, and connections to potential outcomes. Spirtes, Glymour, and Scheines (1993) — working from philosophy at Carnegie Mellon — developed constraint-based algorithms for learning causal structure from data.

Convergence. The three frameworks are closely related and often intertranslatable, but exact equivalence

requires additional assumptions. These notes work mainly in an NPSEM-IE setting, where translation between the languages is especially clean.

1.3 The Core Distinction

The single most important inequality of this course is:

$$\underbrace{P(y \mid x,\, \mathrm{do}(T{=}t))}_{\text{interventional}} \neq \underbrace{P(y \mid x,\, T{=}t)}_{\text{observational}} \qquad \text{whenever } T \text{ is endogenous.} \tag{1.2}$$

Definition: Conditional Exchangeability

Given observed covariates X and confounders U, the treatment assignment is *conditionally exchangeable with the intervention* if:

$$Y(t) \perp\!\!\!\perp T \mid X,\, U, \qquad \text{or equivalently,} \qquad P(y \mid x,\, T{=}t,\, U{=}u) = P(y \mid x,\, \mathrm{do}(T{=}t),\, U{=}u).$$

Once we condition on all common causes (X, U), observing $T{=}t$ and intervening to set $T{=}t$ produce the same distribution of Y.

Definition: Positivity

The treatment assignment satisfies *positivity* if, for every treatment value t of interest, $p_{T|X,U}(t \mid x, u) > 0$ for almost all (x, u). This ensures that $P(y \mid x, T{=}t, U{=}u)$ is well-defined for every treatment value of interest.

Proposition: Back-Door Adjustment Formula

Under conditional exchangeability and positivity:

$$P(y \mid x,\, \mathrm{do}(T{=}t)) = \int P(y \mid x,\, T{=}t,\, U{=}u)\; p(u \mid x)\; du. \tag{1.3}$$

Proof

By positivity, $P(y \mid x, T{=}t, U{=}u)$ is well-defined. Applying the law of total probability to the interventional density:

$$P(y \mid x,\, \mathrm{do}(T{=}t)) = \int P(y \mid x,\, U{=}u,\, \mathrm{do}(T{=}t))\; p(u \mid x)\; du.$$

The weight is $p(u \mid x)$, not $p(u \mid x, T{=}t)$, because under $\mathrm{do}(T{=}t)$ the treatment is set externally and carries no information about U. By conditional exchangeability, $P(y \mid x, U{=}u, \mathrm{do}(T{=}t)) = P(y \mid x, T{=}t, U{=}u)$. Substituting completes the proof. □

Remark: Confounding Discrepancy

What OLS estimates is the observational density $P(y \mid x, T{=}t) = \int P(y \mid x, T{=}t, U{=}u)\; p(u \mid x, T{=}t)\; du$. The kernel $P(y \mid x, T{=}t, U{=}u)$ is the same as in Equation 1.3, but averaged over the *selection-distorted* weight $p(u \mid x, T{=}t)$ rather than the marginal $p(u \mid x)$. The confounding discrepancy:

$$P(y \mid x,\, T{=}t) - P(y \mid x,\, \mathrm{do}(T{=}t)) = \int P(y \mid x,\, T{=}t,\, U{=}u)\big[\,p(u \mid x,\, T{=}t) - p(u \mid x)\,\big]\, du$$

is nonzero when U and T are dependent *and* U affects the conditional distribution of Y.

1.3.1 Two Scenarios: Observed vs. Unobserved Confounder

Scenario 1: U is observed (no unmeasured confounding). When every component of U is recorded, both $P(y \mid x, T{=}t, U{=}u)$ and $p(u \mid x)$ can be estimated, and the back-door formula Equation 1.3 is a functional of the observable distribution. A special case is *unconfoundedness*: if $U \perp\!\!\!\perp T \mid X$ already holds, then $p(u \mid x, T{=}t) = p(u \mid x)$, and the formula reduces to $\mathbb{E}[Y \mid X{=}x, T{=}t]$. This is the key assumption underlying regression adjustment and propensity score methods (Chapters 5 and 6).

Scenario 2: U is unobserved (unmeasured confounding). When U is latent, the back-door formula is not usable from data. The main identification strategies are **instrumental variables** (Chapter 7), the **front-door criterion** (Chapter 3, applied in Chapter 8), and **sensitivity analysis** (Chapter 9).

	Scenario 1: U observed	**Scenario 2: U unobserved**
Confounding type	No unmeasured confounding	Unmeasured confounding
Back-door formula usable?	Yes: both terms estimable	No: U latent
Identification route	Back-door adjustment	IV, front-door, …
Key assumption	$Y(t) \perp\!\!\!\perp T \mid X, U$	IV, front-door, or partial-identification assumptions
Chapters in this course	Chs. 3–6 (back-door); Chs. 10–11 (IPW, DR)	Ch. 7 (IV); Ch. 8 (front-door); Ch. 9 (sensitivity)

1.4 The Gaussian Linear Confounded Model

Example: Gaussian Linear Confounded Model

The structural equations are:

$$Y = \beta T + \gamma U + \varepsilon, \qquad T = \alpha U + \delta, \tag{1.4}$$

where $U \sim \mathcal{N}(0, \sigma_U^2)$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$ are mutually independent. Here β is the causal effect of T on Y, γ is the direct effect of U on Y, and α is the direct effect of U on T.

The endogeneity of T is immediate: $\mathrm{Cov}(T, \gamma U + \varepsilon) = \gamma\alpha\sigma_U^2 \neq 0$ whenever $\alpha \neq 0$ and $\gamma \neq 0$. Consequently, OLS regressing Y on T does not estimate β.

```
\usetikzlibrary{arrows.meta, positioning}
\newcommand{\doop}{\mathrm{do}}
\definecolor{isublue}{RGB}{30,56,100}
\definecolor{defbg}{RGB}{238,244,251}
\definecolor{darkgrey}{RGB}{80,80,80}
\tikzset{
  node/.style={circle,draw=isublue,fill=defbg,thick,minimum size=7mm,font=\small\bfseries},
  unode/.style={circle,draw=darkgrey,fill=gray!10,dashed,thick,minimum size=7mm,font=\small\bfseries
  mnode/.style={rectangle,rounded corners=3pt,draw=darkgrey,fill=gray!15,thick,minimum size=7mm,font=
  edge/.style={-{Stealth[length=5pt]},thick,color=isublue},
  dedge/.style={-{Stealth[length=5pt]},thick,color=darkgrey,dashed}
}
\begin{tikzpicture}[node distance=1.8cm]
  \node[node]  (T1) at (0.5,0)    {$T$};
  \node[node]  (Y1) at (2.5,0)    {$Y$};
  \node[unode] (U1) at (1.5,-1.3) {$U$};
  \draw[edge]  (T1) -- (Y1);
  \draw[dedge] (U1) -- (T1);
  \draw[dedge] (U1) -- (Y1);
  \node[font=\small\itshape,color=darkgrey] at (1.5,-2.1) {Observational DAG};
  \node[mnode] (T2) at (5.5,0)    {$t$};
  \node[node]  (Y2) at (7.5,0)    {$Y$};
  \node[unode] (U2) at (6.5,-1.3) {$U$};
```

```
  \draw[edge]  (T2) -- (Y2);
  \draw[dedge, opacity=0.25] (U2) -- (T2);
  \draw[dedge] (U2) -- (Y2);
  \node[font=\small\itshape,color=darkgrey] at (6.5,-2.1) {Mutilated DAG: $\doop(T{=}t)$};
\end{tikzpicture}
```

Observational DAG *Mutilated DAG:* do(T=t)

Figure 1.2: Graph surgery on the Gaussian confounded model. The faded arrow in the mutilated DAG has been severed by the intervention do(T=t). The back-door path $T \leftarrow U \rightarrow Y$ is eliminated; only the causal path $T \rightarrow Y$ remains active.

1.4.1 Two Explicit Densities

Write $\sigma_T^2 = \alpha^2\sigma_U^2 + \sigma_\delta^2$ for the marginal variance of T.

Interventional density. Set $T = t$ externally; the back-door path through U is severed, so $U \sim \mathcal{N}(0, \sigma_U^2)$ independently:

$$P(y \mid \text{do}(T{=}t)) = \mathcal{N}(\beta t,\ \ \gamma^2\sigma_U^2 + \sigma_\varepsilon^2)\,. \tag{1.5}$$

Observational density. Since (Y, T) is jointly normal, $\text{Cov}(Y, T) = \beta\sigma_T^2 + \gamma\alpha\sigma_U^2$, giving:

$$P(y \mid T{=}t) = \mathcal{N}\left(\left(\beta + \frac{\gamma\alpha\sigma_U^2}{\sigma_T^2}\right)t,\ \ \frac{\gamma^2\sigma_U^2\sigma_\delta^2}{\sigma_T^2} + \sigma_\varepsilon^2\right). \tag{1.6}$$

1.4.2 The Endogeneity Gap

	Interventional	**Observational**
Mean	βt	$\left(\beta + \frac{\gamma\alpha\sigma_U^2}{\sigma_T^2}\right)t$
Variance	$\gamma^2\sigma_U^2 + \sigma_\varepsilon^2$	$\frac{\gamma^2\sigma_U^2\sigma_\delta^2}{\sigma_T^2} + \sigma_\varepsilon^2$
Residual	$(\gamma U{+}\varepsilon) \perp\!\!\!\perp T$	$(\gamma U{+}\varepsilon)$ correlated with T
Identified by	RCT (or IV, Chapter 7)	OLS
Equal when	$\alpha = 0$ or $\gamma = 0$ (no confounding)	

The *endogeneity bias* of OLS is $\gamma\alpha\sigma_U^2/\sigma_T^2$. With $\alpha = \gamma = 1$ and $\sigma_U^2 = \sigma_\delta^2 = 1$: bias $= 1/(1+1) = 0.5$.

1.4.3 Lab: Simulating the Two Densities

The analytical results were verified by simulation using $n = 10{,}000$ draws with parameters $\beta = 2$, $\alpha = 1$, $\gamma = 1$, $\sigma_U = \sigma_\varepsilon = \sigma_\delta = 1$. (R code: `chapter1_lab.R`.)

Experiment 1: OLS bias. By the omitted-variable bias formula, the OLS probability limit is $\beta + \gamma\alpha\sigma_U^2/\sigma_T^2 = 2 + 0.5 = 2.5$.

	Simulated	Theory	Formula
OLS slope	2.5147	2.5000	$\beta + \gamma\alpha\sigma_U^2/\sigma_T^2$
True β	—	2.0000	β
Bias	0.5147	0.5000	$\gamma\alpha\sigma_U^2/\sigma_T^2$

OLS overestimates the causal effect by 25%.

Experiment 2: Two-sample comparison at $t_0 = 1$. The exact observational conditional distribution is $\mathcal{N}(2.5, 1.5)$; the interventional distribution is $\mathcal{N}(2, 2)$.

	Mean (Sim / Theory)	SD (Sim / Theory)
$\mathbb{E}[Y \mid T{=}1]$	2.497 / 2.500	1.229 / $\sqrt{1.5} = 1.225$
$\mathbb{E}[Y \mid \text{do}(T{=}1)]$	2.001 / 2.000	1.428 / $\sqrt{2} = 1.414$

The variance reduction in the observational distribution reflects $\sigma^2_{\text{obs}}/\sigma^2_{\text{int}} = 1.5/2 = 0.75$: conditioning on $T{=}1$ pins down $\alpha U + \delta$, restricting the spread of Y through the correlation with γU.

Experiment 3: Density overlay for varying α. As α increases from 0 to 1.0, the observational density shifts rightward (growing bias) and narrows (reduced variance). The interventional density $\mathcal{N}(2, 2)$ is invariant to α.

1.5 The Two-Step Paradigm

Causal inference is a two-step discipline. The two steps are logically distinct and require different tools.

> Definition: Identification
>
> A causal quantity $P(y \mid \text{do}(T{=}t))$ is *identified* from the observational distribution P if there exists a functional Ψ such that $P(y \mid \text{do}(T{=}t)) = \Psi(P)$, and $\Psi(P)$ depends only on the observable joint distribution.

Step 1: Identification is a purely mathematical question about the causal model: can the interventional density be expressed as a function of the observed-data distribution? It does not depend on sample size. The main tools are: the back-door criterion (Chapter 3, applied in Chapters 4–6), the front-door criterion (Chapter 3), the three rules of the do-calculus (Chapter 3), and IV assumptions (Chapter 7).

Step 2: Estimation. Once $\Psi(P)$ has been identified, the statistical problem is to estimate the functional $\Psi(P)$ from n observations as efficiently as possible. The main tools are efficient influence functions (Part III), IPW (Part III), doubly robust estimators (Part III), and double machine learning (Part III).

1.5.1 Worked Example: Flu Vaccination and Infection (Simpson's Paradox)

A public health agency observes 1,000 individuals. Each person either received a flu vaccine ($T = 1$) or not ($T = 0$), and either became infected ($Y = 1$) or not. Age group X (elderly = E, young = Y) is recorded. Elderly people are both more likely to be vaccinated and more susceptible to infection, so X confounds the T–Y relationship. The causal DAG has paths: $T \to Y$ (causal) and $T \leftarrow X \to Y$ (back-door).

Observed data:

	Vaccinated ($T = 1$)	Unvaccinated ($T = 0$)	Total
Elderly ($X = E$)	360, 30% infected	40, 50% infected	400
Young ($X = Y$)	60, 5% infected	540, 10% infected	600
Total	**420, 26.4% infected**	**580, 12.8% infected**	**1,000**

Simpson's paradox:

Stratum	Vaccinated	Unvaccinated	Difference	Conclusion
Elderly	30%	50%	−20 pp	vaccine helps
Young	5%	10%	−5 pp	vaccine helps
Aggregate	**26.4%**	**12.8%**	+13.6 **pp**	**vaccine harms??**

Within every stratum the vaccine reduces infection. Yet in the aggregate the vaccinated group has a *higher* infection rate, because 86% of vaccinated are elderly vs. 7% of unvaccinated.

Step 1 (Identification). The set $\{X\}$ satisfies the back-door criterion:

$$P(Y{=}1 \mid \mathrm{do}(T{=}t)) = \sum_{x \in \{E,Y\}} P(Y{=}1 \mid T{=}t, X{=}x)\, P(X{=}x).$$

Step 2 (Estimation). With $\hat{P}(X{=}E) = 0.4$ and $\hat{P}(X{=}Y) = 0.6$:

$$\hat{P}(Y{=}1 \mid \mathrm{do}(T{=}1)) = 0.30 \times 0.4 + 0.05 \times 0.6 = 0.15,$$

$$\hat{P}(Y{=}1 \mid \mathrm{do}(T{=}0)) = 0.50 \times 0.4 + 0.10 \times 0.6 = 0.26.$$

The estimated causal risk difference is $0.15 - 0.26 = -0.11$: the vaccine causally reduces infection probability by 11 percentage points.

The Two Steps, Concretely

Step 1 (identification) established — from the DAG alone — that the back-door formula expresses the causal quantity as a functional of observable data. **Step 2 (estimation)** replaced the population probabilities in that formula with sample proportions.
These two steps are logically separate: Step 1 is a mathematical argument about the causal model that does not depend on sample size; Step 2 is a statistical problem conditional on the identification formula. A researcher who skips Step 1 and reports the raw difference conflates the two and draws the wrong conclusion.

1.6 Summary

1. **Interventional vs. observational distribution.** Causal inference answers questions about interventions $P(y \mid \mathrm{do}(T{=}t))$, not about observations $P(y \mid T{=}t)$. In the Gaussian confounded model, the two densities coincide only when $\alpha = 0$ or $\gamma = 0$.
2. **The causal trinity.** The same causal question can be expressed in three languages — SEM (equation surgery), DAG (graph surgery), and potential outcomes ($Y(t)$). In the SEM framework, $Y(t)$ is defined as the solution for Y in the mutilated system.
3. **Endogeneity bias in the Gaussian confounded model.** The coefficient of T in $\mathbb{E}[Y \mid T{=}t]$ differs from the causal coefficient β by $\gamma\alpha\sigma_U^2/\sigma_T^2$. The observational variance $\gamma^2\sigma_U^2\sigma_\delta^2/\sigma_T^2 + \sigma_\varepsilon^2$ is smaller than the interventional variance $\gamma^2\sigma_U^2 + \sigma_\varepsilon^2$.
4. **Two-step paradigm.** Causal inference is: *identification* (expressing $\Psi(P)$ as a functional of observable data — a mathematical question) followed by *estimation* (constructing $\hat{\Psi}_n$ efficiently — a statistical question). The two steps are analytically separate.

Causal inference is the study of when the observational distribution contains enough information to recover the interventional distribution.

1.7 Problems

1. Do-notation fundamentals. For each statement below, decide whether it refers to an interventional or an observational distribution, and rewrite it unambiguously using do-notation where appropriate.

(a) "The probability that a patient recovers given that they took the drug."
(b) "The probability that a patient would recover if we prescribed the drug to everyone."
(c) "Among students who attended tutoring, the average exam score was 85."
(d) "If we enrolled all students in tutoring, the average exam score would be 85."

2. Deriving the observational density. In the Gaussian confounded model (Equation 1.4), verify Equation 1.6 by carrying out the following steps. Let $\sigma_T^2 = \alpha^2\sigma_U^2 + \sigma_\delta^2$.

(a) Show that (Y, T) is jointly normal by writing both as linear combinations of the independent normals (U, ε, δ).
(b) Compute $\mathrm{Cov}(Y, T)$ and $\mathrm{Var}(T)$.
(c) Apply the conditional normal formula to derive $\mathbb{E}[Y \mid T{=}t]$ and $\mathrm{Var}[Y \mid T{=}t]$, and hence verify Equation 1.6.
(d) Show that the OLS probability limit is $\beta + \gamma\alpha\sigma_U^2/\sigma_T^2$, and interpret this as an omitted-variable bias formula.

3. The causal trinity. Consider the following verbal causal claim: "Aspirin (T) reduces the risk of heart attack (Y) because it inhibits platelet aggregation (M), but patients with pre-existing cardiovascular disease (U, unobserved) are both more likely to take aspirin and more likely to have a heart attack."

(a) Draw the causal DAG implied by this description.
(b) Write down the recursive SEM with appropriate structural functions f_T, f_M, f_Y.
(c) State the SUTVA assumption and discuss whether it is plausible in this setting.
(d) Explain why $\mathbb{E}[Y \mid T{=}1] - \mathbb{E}[Y \mid T{=}0]$ does not identify the causal effect $\mathbb{E}[Y \mid \mathrm{do}(T{=}1)] - \mathbb{E}[Y \mid \mathrm{do}(T{=}0)]$ in this DAG.

Chapter 2

DAGs and d-Separation

Learning Objectives

By the end of this chapter, students should be able to:

1. Construct a DAG from a verbal description of a causal model and identify parents, descendants, ancestors, and non-descendants of any node.
2. Classify every intermediate node on a path as a chain, fork, or collider, and apply the blocking rule for each.
3. Apply the d-separation criterion to determine all conditional independence relationships implied by a DAG.
4. Recognize and avoid collider bias, including the case where a descendant of a collider is conditioned upon.
5. State the Markov factorization and use it to write the joint density of a DAG in terms of local conditional densities.
6. Interpret d-separation results as conditional independence assumptions on observed variables, and recognize their role as the graphical expression of causal assumptions.

Readers who want a gentler introduction to conditional independence, the three basic path motifs, and the Markov-property interpretation may consult Appendix A before or alongside this chapter.

2.1 Directed Acyclic Graphs

DAGs allow us to translate qualitative causal assumptions into quantitative statistical restrictions. Once we draw a graph encoding our causal assumptions, d-separation tells us which conditional independences are implied by the graph and helps identify candidate adjustment sets for removing confounding.

Remark

In this chapter, the word *graph* does not mean a plot of data or a graph of a function. It means a collection of nodes and edges used to represent relationships among variables. Our objective is to understand how such graphs encode statistical structure — especially conditional independence, confounding, and path blocking.

2.1.1 Basic Definition

Definition: Directed Acyclic Graph

A *directed acyclic graph* (DAG) $\mathcal{G} = (\mathcal{V}, E)$ consists of a finite set of nodes $\mathcal{V}$ and a set of directed edges $E \subseteq \mathcal{V} \times \mathcal{V}$ such that there is no directed cycle. Each node represents a variable and each directed edge $A \to B$ encodes that A is a direct cause of B relative to the variables included in the graph.

The acyclicity condition guarantees a topological ordering so that the joint density admits the recursive factorization $p(v_1, \dots, v_k) = \prod_{i=1}^{k} p(v_i \mid \mathrm{Pa}(v_i))$.

2.1.2 Structural Relationships

Definition: Structural Relationships in a DAG

For a node $v \in \mathcal{V}$ in $\mathcal{G}$:

- $\mathrm{Pa}(v) = \{w \in \mathcal{V} : w \to v \in E\}$ are the *parents* of v.
- $\mathrm{De}(v) =$ all nodes reachable from v by directed paths are the *descendants* of v.
- $\mathrm{An}(v) =$ all nodes with a directed path to v are the *ancestors* of v.
- $\mathrm{Nd}(v) = \mathcal{V} \setminus (\mathrm{De}(v) \cup \{v\})$ are the *non-descendants* of v.

Note that $\mathrm{Nd}(v)$ includes the parents of v.

Quick Terminology Summary

If $A \to B$, then A is a *parent* of B and B is a *child* of A. Two nodes are *adjacent* if connected by an edge. A *path* is any sequence of connected nodes regardless of arrow direction; a *directed path* has arrows all pointing in the same direction. A *cycle* is a directed path that returns to its starting node.

Example: Education, Earnings, and Family Background

Education (E) affects earnings (Y); family background (B) is a common cause of both; neighborhood (N) affects education but has no direct effect on earnings. The Markov factorization is $p(n, b, e, y) = p(n)\, p(b)\, p(e \mid n, b)\, p(y \mid e, b)$. Reading off: $\mathrm{Pa}(E) = \{N, B\}$, $\mathrm{Pa}(Y) = \{E, B\}$, $\mathrm{Pa}(N) = \mathrm{Pa}(B) = \emptyset$.

Example: The IV DAG

The canonical IV model involves Z (instrument), T (treatment), Y (outcome), U (unobserved confounder), with $\mathrm{Pa}(T) = \{Z, U\}$, $\mathrm{Pa}(Y) = \{T, U\}$, $\mathrm{Pa}(Z) = \mathrm{Pa}(U) = \emptyset$. The descendants of Z are $\{T, Y\}$; U and Z are non-descendants of each other.

Example: The Mediation DAG

Treatment T affects outcome Y both directly and through mediator M. An unobserved confounder U creates a back-door path $T \leftarrow U \to Y$. $\mathrm{Pa}(T) = \{U\}$, $\mathrm{Pa}(M) = \{T\}$, $\mathrm{Pa}(Y) = \{M, T, U\}$. The total effect of T on Y travels along two paths: $T \to Y$ (direct) and $T \to M \to Y$ (indirect). Note: this is a standard mediation DAG, not a front-door DAG. For the front-door criterion, every directed path from T to Y must pass through M.

2.2 Paths, Blocking, and d-Separation

A DAG encodes direct causal relationships through its edges, but variables can also be statistically related through longer routes. The concept of a *path* formalizes these routes, and d-separation answers whether a path transmits or blocks statistical dependence.

2.2.1 Path Structures

Definition: Path

A *path* between nodes X and Y in $\mathcal{G}$ is any sequence of distinct nodes $X = V_0, V_1, \ldots, V_k = Y$ such that each consecutive pair is connected by an edge in either direction.

Every intermediate node V_m on a path plays one of three structural roles. In a **chain** ($V_{m-1} \to V_m \to V_{m+1}$), V_m is a causal intermediary. In a **fork** ($V_{m-1} \leftarrow V_m \to V_{m+1}$), V_m is a common cause. In a **collider** ($V_{m-1} \to V_m \leftarrow V_{m+1}$), both arrows point *into* V_m; this asymmetric case has the opposite behavior under conditioning from the other two.

2.2.2 Blocking Rules

Structure	**Pattern**	**Effect of conditioning**	**Key intuition**
Chain	$A \to M \to B$	Blocks the path	Causal transmission
Fork	$A \leftarrow M \to B$	Blocks the path	Common cause
Collider	$A \to M \leftarrow B$	**Opens** the path	Common effect

The critical asymmetry: colliders are *closed by default* and *opened* by conditioning, while chains and forks are *open by default* and *closed* by conditioning.

Three-Node Motifs at a Glance

In a *chain* $A \to M \to B$: conditioning on M blocks the path. In a *fork* $A \leftarrow M \to B$: conditioning on M removes the spurious association. In a *collider* $A \to M \leftarrow B$: the path is blocked by default, but conditioning on M — or any descendant of M — opens it. These three motifs are the local building blocks of d-separation.

Chain: Smoking → Tar → Cancer. Marginally, smokers have elevated cancer rates (path is open). Conditioning on tar level blocks the path: once tar is fixed, the causal channel is fully accounted for.

Fork: Poverty → Poor Diet; Poverty → Lack of Exercise. Unconditionally, diet quality and exercise are correlated through poverty. Conditioning on income level removes the spurious association.

Collider: Accident → Hospitalization ← Cancer. In the general population, $A \perp\!\!\!\perp B$. Among hospitalized patients — conditioning on M — the two become negatively associated. This is Berkson's bias (Berkson 1946).

2.2.3 The d-Separation Criterion

Definition: d-Blocking and d-Separation

A path π is *d-blocked* by a set $\mathbf{S}$ if:

1. π contains a chain $A \to M \to B$ or fork $A \leftarrow M \to B$ with $M \in \mathbf{S}$; or
2. π contains a collider $A \to C \leftarrow B$ with $C \notin \mathbf{S}$ and no descendant of C is in $\mathbf{S}$.

Nodes X and Y are *d-separated* by $\mathbf{S}$, written $(X \perp\!\!\!\perp Y \mid \mathbf{S})_{\mathcal{G}}$, if every path between X and Y in $\mathcal{G}$ is d-blocked by $\mathbf{S}$.

Example: Applying the Criterion to the Three Toy Settings

(i) Chain. $A \to M \to B$: $\mathbf{S} = \emptyset$ gives $A \not\perp\!\!\!\perp B$; $\mathbf{S} = \{M\}$ gives $A \perp\!\!\!\perp B \mid M$.
(ii) Fork. $A \leftarrow M \to B$: $\mathbf{S} = \emptyset$ gives $A \not\perp\!\!\!\perp B$; $\mathbf{S} = \{M\}$ gives $A \perp\!\!\!\perp B \mid M$.
(iii) Collider. $A \to M \leftarrow B$: $\mathbf{S} = \emptyset$ gives $A \perp\!\!\!\perp B$ (collider blocks the path); $\mathbf{S} = \{M\}$ gives $A \not\perp\!\!\!\perp B \mid M$ (conditioning opens the path — Berkson's bias).

Example: A Direct Probability Proof for the Chain

Consider the chain $X_1 \to X_2 \to X_3$ with factorization $p(x_1, x_2, x_3) = p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_2)$. Conditioning on X_2:

$$p(x_1, x_3 \mid x_2) = \frac{p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_2)}{p(x_2)} = p(x_1 \mid x_2)\,p(x_3 \mid x_2).$$

Hence $X_1 \perp\!\!\!\perp X_3 \mid X_2$. The analogous fork case $X_1 \leftarrow X_2 \to X_3$ is left as an exercise.

Theorem: Soundness of d-Separation [@pearl2009causality, Theorem 1.2.4]

If $(X \perp\!\!\!\perp Y \mid \mathbf{S})_{\mathcal{G}}$, then $X \perp\!\!\!\perp Y \mid \mathbf{S}$ in every distribution that is Markov with respect to $\mathcal{G}$.

Remark: Soundness, Not Converse

The theorem states that d-separation *implies* conditional independence for every distribution Markov with respect to $\mathcal{G}$. The converse need not hold without additional assumptions such as *faithfulness*: a conditional independence may occur because of special parameter values even when the corresponding nodes are not d-separated in the graph.

2.2.4 Practical Ways to Check d-Separation

1. Bayes-Ball Intuition. Imagine releasing a ball from X asking whether it can reach Y, with nodes in $\mathbf{S}$ shaded. The ball: passes through chain/fork nodes not in $\mathbf{S}$; stops at chain/fork nodes in $\mathbf{S}$; stops at colliders not in $\mathbf{S}$ (with no descendant in $\mathbf{S}$); passes through colliders in $\mathbf{S}$ (or with a descendant in $\mathbf{S}$). The algorithm is formalized in Shachter (1998).

2. Moral Graph Transformation. (1) Restrict to the ancestral set $\mathrm{An}(X \cup Y \cup \mathbf{S})$. (2) Moralize: for every collider $A \to C \leftarrow B$, add an undirected edge $A - B$. (3) Remove all edge directions. (4) Delete all nodes in $\mathbf{S}$. If X and Y are disconnected, then $(X \perp\!\!\!\perp Y \mid \mathbf{S})_{\mathcal{G}}$. See Lauritzen (1996, Ch. 3) for a full treatment.

Example: Checking d-Separation in a Four-Node DAG

Consider the DAG with edges $Z \to T$, $U \to T$, $T \to Y$, $U \to Y$.
Query 1. Is $(Z \perp\!\!\!\perp U)_{\mathcal{G}}$?
Bayes-ball: The only path is $Z \to T \leftarrow U$. Node T is a collider with $T \notin \emptyset$, so the ball stops. Hence $(Z \perp\!\!\!\perp U)_{\mathcal{G}}$.
Moral graph: Ancestral set of $\{Z, U\}$ is $\{Z, U\}$ (neither has parents); T and Y are discarded. No edges, no colliders. Z and U are disconnected $\Rightarrow (Z \perp\!\!\!\perp U)_{\mathcal{G}}$.
Query 2. Is $(Z \perp\!\!\!\perp U \mid T)_{\mathcal{G}}$?
Bayes-ball: Same path $Z \to T \leftarrow U$. Now $T \in \{T\}$: the collider is observed, so the ball passes through. Hence $Z \not\perp\!\!\!\perp U \mid T$.
Moral graph: Ancestral set of $\{Z, U, T\}$ includes T. Moralize: add edge $Z - U$ for the collider $Z \to T \leftarrow U$. After removing T, the remaining graph has edge $Z - U$. Connected $\Rightarrow Z \not\perp\!\!\!\perp U \mid T$.
Comparing: Z and U are marginally independent (instrument is exogenous) but become dependent once we condition on T — Berkson's bias.

Example: Practice DAG — Eight-Node Graph

Consider the DAG with nodes W, X_1, T, M_1, M_2, Y, C, D and edges $W \to X_1$, $W \to T$, $W \to Y$, $T \to M_1$, $M_1 \to Y$, $M_1 \to M_2$, $X_1 \to C$, $M_2 \to C$, $C \to D$.
(1) Is $X_1 \perp\!\!\!\perp Y$? No. Two open paths: $X_1 \leftarrow W \to Y$ (fork at W, unblocked) and $X_1 \leftarrow W \to T \to M_1 \to Y$ (also open). The path $X_1 \to C \leftarrow M_2$ ends at a collider C with $C \notin \emptyset$, so it is blocked.
(2) Is $X_1 \perp\!\!\!\perp Y \mid W$? Yes. Both open paths above pass through the fork W; conditioning on W

blocks them. The remaining path $X_1 \to C \leftarrow M_2 \leftarrow \cdots$ has collider $C \notin \{W\}$ — blocked. All paths blocked.
(3) Is $T \perp\!\!\!\perp M_2 \mid M_1$? Yes. The direct path $T \to M_1 \to M_2$ is a chain with $M_1 \in \{M_1\}$ — blocked. Other paths through W or Y also contain blocked colliders. All paths blocked.
(4) Does conditioning on C create collider bias? Yes. C is a collider on $X_1 \to C \leftarrow M_2$. Marginally $X_1 \perp\!\!\!\perp M_2$ (blocked by the collider). Conditioning on C opens the path, inducing spurious association.
(5) Does conditioning on D (a descendant of C) open the path $X_1 \to C \leftarrow M_2$? Yes. By clause (2) of the d-blocking definition, a collider path is unblocked whenever the collider *or any of its descendants* is in the conditioning set.

2.3 Collider Bias and the IV DAG

2.3.1 Berkson's Bias

Definition: Collider Bias

Collider bias is the spurious association induced between two variables when one conditions on their common effect (a collider), or on a descendant of that collider. In a path $A \to C \leftarrow B$, conditioning on C or any descendant of C can make A and B statistically dependent even if they are marginally independent.

Collider Bias vs. Confounding

Unlike confounding, which is *removed* by conditioning on a common cause, collider bias is *created* by conditioning on a common effect. Adding a collider or a descendant of a collider to the adjustment set introduces bias rather than reducing it.

Example: Collider Bias through Selection — Talent and Wealth

Suppose both talent (A) and wealth (B) increase the probability of admission to an elite school, so admission (S) is a collider: $A \to S \leftarrow B$. In the general population, $A \perp\!\!\!\perp B$ (the collider blocks the path). Among admitted students — conditioning on S — lower talent makes higher wealth more likely, and vice versa. This is collider bias: restricting to admitted students is precisely conditioning on the common effect.

2.3.2 Full d-Separation Analysis of the IV DAG

We work through the IV DAG with edges $Z \to T$, $T \to Y$, $U \to T$, $U \to Y$, with U *unobserved*. Because U cannot be conditioned on, the back-door path $T \leftarrow U \to Y$ cannot be blocked by adjustment, making the instrument Z the only route to identification.

Classifying paths by type:

Endpoints	Path	Type	Why
$Z \leftrightarrow Y$	$Z \to T \to Y$	*Causal*	Directed; carries the IV signal
$Z \leftrightarrow Y$	$Z \to T \leftarrow U \to Y$	*Associational*	Collider at T; activates back-door via U
$Z \leftrightarrow U$	$Z \to T \leftarrow U$	*Associational*	Collider at T; dormant unless T conditioned on

Question 1. Is $Z \perp\!\!\!\perp Y$? No. The directed path $Z \to T \to Y$ is open (chain, not conditioned on T). The path $Z \to T \leftarrow U \to Y$ has a collider at T with $T \notin \emptyset$ — blocked.

Question 2. Is $Z \perp\!\!\!\perp Y \mid T$? No. The causal path $Z \to T \to Y$ is blocked (conditioning on T in the chain). But the collider at T in $Z \to T \leftarrow U \to Y$ is now opened. Hence $Z \not\perp\!\!\!\perp Y \mid T$.

Question 3. Is $Z \perp\!\!\!\perp Y \mid \{T, U\}$? Yes. The causal path is blocked by conditioning on T. The path through U has the collider at T opened (by conditioning on T), but then blocked at U (by conditioning on U). Both paths blocked.

Question 4. Is $Z \perp\!\!\!\perp U$? Yes. The only path is $Z \to T \leftarrow U$ with a collider at T. Since $T \notin \emptyset$ and no descendant of T is in $\emptyset$, the path is blocked.

Interpretation. Question 1 establishes *relevance*: Z has an open causal path to Y through T. Question 4 establishes *exogeneity*: the instrument is independent of unmeasured confounding. Question 3 establishes the *exclusion restriction*: once T and U are held fixed, Z carries no residual information about Y. Note that this requires observing U, which is unavailable by assumption; the IV strategy exploits the exclusion restriction indirectly through the moment condition $\mathbb{E}[\varepsilon \cdot Z] = 0$ (Chapter 7). Question 2 is the *warning*: conditioning on T alone simultaneously closes the causal channel and opens the confounded channel — worse than no adjustment at all.

The IV DAG encodes the substantive assumptions behind instrument validity graphically, but does not make them testable in any strong sense.

2.4 The Markov Property and Factorization

Up to this point, we have used DAGs qualitatively. We now connect the graph to probability algebra: the same parent structure that governs d-separation also determines how the joint distribution factorizes.

Without structural assumptions, any joint distribution can always be written by repeated conditioning, but that generic representation is too high-dimensional to reveal much structure. A DAG becomes statistically meaningful because, together with the Markov property, it replaces the generic factorization by a sparse one involving only the parents of each node.

Proposition: Conditional Independence in the Three-Node Chain

Let $X_1 \to X_2 \to X_3$ be a chain with Markov factorization $p(x_1, x_2, x_3) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_2)$. Then $X_1 \perp\!\!\!\perp X_3 \mid X_2$. (Proved in the chain example above; the fork case is analogous.)

Definition: Local Markov Property

A distribution P satisfies the *local Markov property* with respect to $\mathcal{G}$ if, for every node $V_i \in \mathcal{V}$:

$$V_i \perp\!\!\!\perp \left[\mathrm{Nd}(V_i) \setminus \mathrm{Pa}(V_i)\right] \,\Big|\, \mathrm{Pa}(V_i).$$

Once we condition on the direct causes of a node, that node is independent of all variables that are neither its descendants nor its parents.

Remark

The *global Markov property* is the statement that every d-separation in $\mathcal{G}$ implies a conditional independence in P — precisely the content of the Soundness theorem. For DAGs, the local and global Markov properties are equivalent (Lauritzen 1996, Proposition 3.27): the local property implies the global one via the Markov factorization.

An immediate consequence is the **Markov factorization**:

$$p(v_1, \ldots, v_k) = \prod_{i=1}^{k} p(v_i \mid \mathrm{Pa}(v_i))\,. \tag{2.1}$$

Example: Markov Factorization in the Education DAG

For the DAG $N \to E$, $B \to E$, $B \to Y$, $E \to Y$: $p(n, b, e, y) = p(n)\,p(b)\,p(e \mid n, b)\,p(y \mid e, b)$. A regression of Y on E alone is confounded by B. Conditioning on B blocks the back-door path $E \leftarrow B \to Y$ and identifies $P(y \mid \mathrm{do}(E{=}e))$ via the back-door formula (Chapter 3).

Example: Markov Factorization in the IV DAG

Including the latent U: $p(z, t, y, u) = p(z)\,p(u)\,p(t \mid z, u)\,p(y \mid t, u)$. The observable factorization is obtained by marginalizing over U:

$$p(z, t, y) = \int p(z)\,p(u)\,p(t \mid z, u)\,p(y \mid t, u)\,\mathrm{d}u.$$

This cannot be simplified to $p(z)\,p(t \mid z)\,p(y \mid t)$ when U is a common cause of T and Y — this is precisely the endogeneity problem.

2.5 Worked Example: The Education–Earnings DAG

This section is the template for how we will use DAGs throughout the course: specify the DAG, read off the factorization, use d-separation to identify the implied conditional independences, and interpret in causal terms.

Practice. Before working through the example below, return to the eight-node DAG example and rework each of the five queries from scratch, following three steps: (1) list every path; (2) classify every intermediate node as chain, fork, or collider; (3) determine which paths are blocked or open.

The causal story. Education (E) affects earnings (Y); family background (B) is a common cause of both; neighborhood (N) affects education but has no direct effect on earnings. All four variables are observed (no latent confounders).

Step 1 — Markov factorization. $\mathrm{Pa}(N) = \mathrm{Pa}(B) = \emptyset$, $\mathrm{Pa}(E) = \{N, B\}$, $\mathrm{Pa}(Y) = \{E, B\}$:

$$p(n, b, e, y) = p(n)\,p(b)\,p(e \mid n, b)\,p(y \mid e, b).$$

Step 2 — d-Separation. There are two paths between N and Y:

- Path 1: $N \to E \to Y$ (a chain through E).
- Path 2: $N \to E \leftarrow B \to Y$ (a collider at E, followed by the fork leg $B \to Y$).

(i) Is $(N \perp\!\!\!\perp Y)_{\mathcal{G}}$*?* Path 1 is a chain with nothing conditioned on, so it is open. Path 2 has a collider at E (not conditioned on), so it is blocked. One open path: $N \not\perp\!\!\!\perp Y$.

(ii) Is $(N \perp\!\!\!\perp Y \mid E)_{\mathcal{G}}$*?* Path 1 is blocked (conditioning on E in the chain). Path 2 has a collider at E opened by conditioning, and the activated path continues through B (not conditioned on), so it is open. Hence $N \not\perp\!\!\!\perp Y \mid E$ — a collider trap: conditioning on E introduces bias, not removes it.

(iii) Is $(N \perp\!\!\!\perp Y \mid \{E, B\})_{\mathcal{G}}$*?* Path 1 is blocked by E. Path 2 is opened at collider E but then blocked at B. All paths blocked: $(N \perp\!\!\!\perp Y \mid E, B)_{\mathcal{G}}$.

(iv) Is $(N \perp\!\!\!\perp B)_{\mathcal{G}}$*?* The only path $N \to E \leftarrow B$ has a collider at E (not conditioned on). Path is blocked: $N \perp\!\!\!\perp B$.

Step 3 — Conditional Independence. By the Soundness theorem:

$$N \perp\!\!\!\perp B, \qquad N \perp\!\!\!\perp Y \mid E, B, \qquad N \not\perp\!\!\!\perp Y, \qquad N \not\perp\!\!\!\perp Y \mid E.$$

The statement $N \not\perp\!\!\!\perp Y \mid E$ is an important warning: conditioning only on education when studying the neighborhood–earnings association *introduces* bias through the activated collider path.

Step 4 — Identification. We wish to identify $P(y \mid \mathrm{do}(E{=}e))$. The only back-door path is $E \leftarrow B \to Y$ (a fork at B).

- $\mathbf{S} = \{B\}$: blocks the fork $E \leftarrow B \rightarrow Y$, and B is not a descendant of E. Valid.
- $\mathbf{S} = \{N\}$: does not block $E \leftarrow B \rightarrow Y$. Invalid.

With $\mathbf{S} = \{B\}$, the back-door formula (Chapter 3) gives:

$$P(y \mid \text{do}(E{=}e)) = \sum_b P(y \mid e, b)\, P(b),$$

identifying the causal effect entirely from observational data.

2.6 The Big Picture

The central logic runs from a qualitative causal graph to an identification formula:

structural assumptions $\Longrightarrow$ DAG $\Longrightarrow$ d-separation relations $\Longrightarrow$ conditional independences (Markov) $\Longrightarrow$ identification for

Chapter 2 establishes the first four links in this chain. Chapters 3 and beyond use the same graphical machinery to derive specific identification results: back-door adjustment, front-door adjustment, and do-calculus formulas.

2.7 Summary

1. **DAGs as causal structure.** A DAG encodes which variables are causally connected, which are potential confounders, and which variables may block or open paths.
2. **Three-node motifs.** Every path is built from chains, forks, and colliders. Chains and forks are open by default and are blocked by conditioning on the middle node. Colliders are blocked by default and are opened by conditioning on the collider or on one of its descendants.
3. **d-Separation and conditional independence.** X and Y are d-separated by $\mathbf{S}$ exactly when every path between them is blocked by $\mathbf{S}$. Under the Markov property, d-separation implies the corresponding conditional independence. Some implications may be assessed empirically, but assumptions involving unobserved variables are generally not testable.
4. **Markov factorization.** The local Markov property yields $p(v_1, \ldots, v_k) = \prod_{i=1}^{k} p(v_i \mid \text{Pa}(v_i))$, the bridge from graphical structure to probability calculus.
5. **Collider bias.** Conditioning on a collider — or on a descendant of a collider — can induce a spurious association between otherwise independent variables (Berkson 1946). This is the opposite of confounding adjustment.
6. **A practical workflow.** To analyze a DAG: identify the graph structure, determine which paths are open or blocked, translate d-separation statements into conditional independences under the Markov property, and interpret in light of the causal question.

2.8 Problems

1. Warm-up: a single collider. Consider the DAG $X \rightarrow Y \leftarrow Z$.

(a) Identify the structural role of Y on the path $X \rightarrow Y \leftarrow Z$.
(b) Is $X \perp\!\!\!\perp Z$? Apply the d-separation criterion and state which blocking rule applies.
(c) Is $X \perp\!\!\!\perp Z \mid Y$? Explain what happens to the path when Y is conditioned on, and describe in one sentence the real-world phenomenon this illustrates.

2. d-Separation practice. Consider the DAG: $A \rightarrow B \rightarrow D$, $A \rightarrow C \rightarrow D$, $B \rightarrow E$, $C \rightarrow E$.

(a) List all paths between A and E. (*Hint*: there are four paths; two pass through D.)
(b) For each path, identify the role (chain, fork, collider) of each intermediate node.
(c) Does $\{B, C\}$ d-separate A and E?
(d) Does $\{D\}$ d-separate B and C? What type of node is D on the path $B \rightarrow D \leftarrow C$?

3. Berkson's bias. Suppose X and Y are independent standard normal variables, and let $S = \mathbf{1}[X + Y > 0]$.

(a) Verify analytically that $\text{Cov}(X, Y \mid S{=}1) < 0$.
(b) Draw the DAG for (X, Y, S) and identify S as a collider.
(c) Explain in one sentence why restricting the analysis to the subsample with $S = 1$ biases estimates of any association between X and Y.

4. Markov factorization and collider activation. Consider the DAG with edges $A \to E$, $A \to W$, $F \to E$, $E \to W$, where A = ability, F = family income, E = education, W = wages.

(a) Write down the Markov factorization $p(a, f, e, w)$.
(b) Is $(F \perp\!\!\!\perp W)_{\mathcal{G}}$? List all paths between F and W and determine which are open.
(c) Is $(F \perp\!\!\!\perp W \mid E)_{\mathcal{G}}$? Identify the role of E on each path.
(d) A researcher regresses W on E and F, omitting A. Is the coefficient on E a causal effect? Explain using the graph.

5. Soundness theorem for the collider. Consider the collider $A \to M \leftarrow B$ with factorization $p(a, m, b) = p(a)\, p(b)\, p(m \mid a, b)$.

(a) By marginalizing over M, show that $A \perp\!\!\!\perp B$ in the joint distribution. This verifies the Soundness theorem for $\mathbf{S} = \emptyset$.
(b) Show that conditioning on M breaks this independence: write out $p(a, b \mid m)$ and explain why it does *not* factorize into $p(a \mid m)\, p(b \mid m)$ in general.
(c) Explain in one sentence why parts (a) and (b) together are consistent with the Soundness theorem. (*Hint*: the theorem is a one-directional statement.)

6. Terminology check. Consider the DAG with edges $U \to X$, $U \to Z$, $X \to W$, $Z \to W$, $W \to Y$.

(a) Identify the parents, children, ancestors, descendants, and non-descendants of node W.
(b) List all pairs of adjacent nodes.
(c) Which pairs of nodes are connected by a directed path? List every such pair and the corresponding path.
(d) Write the Markov factorization $p(u, x, z, w, y)$.

7. Markov factorization and local Markov property. Consider the DAG with edges $X_1 \to X_2$, $X_1 \to X_3$, $X_2 \to X_4$, $X_3 \to X_4$.

(a) Write the joint density $p(x_1, x_2, x_3, x_4)$ implied by the Markov factorization.
(b) State the local Markov property for each of the four nodes.
(c) Is $(X_2 \perp\!\!\!\perp X_3)_{\mathcal{G}}$? Is $(X_2 \perp\!\!\!\perp X_3 \mid X_1)_{\mathcal{G}}$? Justify each answer by listing all paths.

8. Toy proof: conditional independence in the fork. Consider the fork $X_1 \leftarrow X_2 \to X_3$ with factorization $p(x_1, x_2, x_3) = p(x_2)\, p(x_1 \mid x_2)\, p(x_3 \mid x_2)$.

(a) Show directly, by conditioning on $X_2 = x_2$, that $X_1 \perp\!\!\!\perp X_3 \mid X_2$.
(b) Is $X_1 \perp\!\!\!\perp X_3$ marginally? Justify both graphically and algebraically.
(c) Explain in one sentence what the fork represents substantively and why conditioning on the common cause removes the association.

9. d-Separation in the practice DAG. Refer to the eight-node DAG in the Practice DAG example (nodes W, X_1, T, M_1, M_2, Y, C, D; edges $W \to X_1$, $W \to T$, $W \to Y$, $T \to M_1$, $M_1 \to Y$, $M_1 \to M_2$, $X_1 \to C$, $M_2 \to C$, $C \to D$). For each query, state whether it is true or false and justify by listing all relevant paths.

(a) $X_1 \perp\!\!\!\perp Y$
(b) $X_1 \perp\!\!\!\perp Y \mid W$
(c) $T \perp\!\!\!\perp M_2 \mid M_1$
(d) Does conditioning on C induce collider bias between X_1 and M_2? Identify the specific path opened.
(e) Does conditioning on D (a descendant of C) open the path $X_1 \to C \leftarrow M_2$? State which clause of the d-blocking definition applies.

Chapter 3

The Do-Calculus and Identification Criteria

Learning Objectives

By the end of this chapter, students should be able to:

1. Define the two intervention graph operations $\mathcal{G}_{\overline{X}}$ and $\mathcal{G}_{\underline{X}}$, explain the causal meaning of $\mathcal{G}_{\overline{X}}$, and describe the technical role of $\mathcal{G}_{\underline{X}}$.
2. Apply the back-door criterion to determine whether a set $\mathbf{S}$ identifies a causal effect, and write the back-door adjustment formula.
3. Apply the front-door criterion in graphs where the confounder is unobserved but a suitable mediator exists.
4. State all three rules of do-calculus, identify the graph condition that licenses each rule, and use them to prove the back-door and front-door formulas.
5. State the completeness theorem of Shpitser and Pearl (2006) and explain its practical implication: if the do-calculus cannot identify a quantity from the observational distribution relative to the assumed graph, then no purely observational method can do so without additional assumptions or new data.

How to Read This Chapter

This chapter has two pedagogical layers. Section 3.1–Section 8.8 develop the two most important concrete identification criteria — back-door and front-door adjustment — using intervention graphs and direct graphical reasoning. These sections are the core material for a first reading.
Section 3.4–Section 3.6 then place these criteria inside the more general theory of the do-calculus and identifiability. These later sections are conceptually important but more abstract and may be read as a second pass after the concrete criteria are understood.

3.1 From d-Separation to Intervention: Intervention Graphs

In Chapter 2 we learned to read conditional independence from a DAG using d-separation. This chapter takes the next step: translating the graphical language into *identification formulas* — expressions that write the interventional distribution $P(y \mid \mathrm{do}(T{=}t))$ entirely in terms of quantities observable from data.

3.1.1 What Does Intervention Mean?

Conditioning versus intervening. The conditional distribution $P(Y \mid T{=}t)$ describes the subpopulation of units for whom T was *observed* to equal t. The interventional distribution $P(Y \mid \mathrm{do}(T{=}t))$, by contrast, describes the population that would result if T were *externally set* to t for everyone. The two coincide only when T has no unmeasured common causes with Y.

A simple illustration: temperature Z causes both ice-cream sales T and crime Y. $P(Y \mid T{=}t)$ is the crime rate on days when sales happen to equal t — confounded by Z. $P(Y \mid \mathrm{do}(T{=}t))$ is the crime rate we would observe if we *fixed* sales at level t by decree. If T has no direct causal path to Y, then $P(Y \mid \mathrm{do}(T{=}t)) = P(Y)$ for all t.

The structural basis of the do-operator. In the SEM framework, every variable is generated by a structural equation. For the treatment node: $T = f_T(\mathrm{Pa}(T), U_T)$. An intervention $\mathrm{do}(T{=}t)$ *replaces this entire equation* with the constant $T = t$, with two consequences: (1) T is no longer influenced by its parents; (2) all other structural equations remain unchanged.

Graph surgery as the graphical realization. Because $\mathrm{Pa}(T)$ no longer affects T after the intervention, every arrow pointing into T in the DAG should be removed. The resulting graph — the *intervention graph* $\mathcal{G}_{\overline{T}}$ — represents the post-intervention world. D-separation in $\mathcal{G}_{\overline{T}}$ encodes conditional independence in $P(\cdot \mid \mathrm{do}(T{=}t))$, not in the original P. This is the key link that allows graphical reasoning to answer causal questions.

3.1.2 The Two Graph Operations

Definition: Intervention Graphs $\mathcal{G}_{\overline{X}}$ and $\mathcal{G}_{\underline{X}}$

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a DAG and $X \subseteq \mathcal{V}$.

- $\mathcal{G}_{\overline{X}}$ (*intervention graph*, or *mutilated graph*) is obtained by **deleting all arrows into** X. This represents $\mathrm{do}(X{=}x)$, severing X's dependence on its former parents.
- $\mathcal{G}_{\underline{X}}$ (*auxiliary observation graph*) is obtained by **deleting all arrows out of** X. This is a *technical device* used in graphical conditions for certain do-calculus steps. It does *not* represent conditioning on X.

3.2 The Back-Door Criterion

Definition: Back-Door Criterion [@pearl1993comment]

A set $\mathbf{S}$ of observed variables satisfies the **back-door criterion** for the effect of T on Y in $\mathcal{G}$ if:

1. No node in $\mathbf{S}$ is a descendant of T.
2. $\mathbf{S}$ blocks every *back-door path* from T to Y — every path that begins with an arrow pointing *into* T.

Remark

The term "back-door" reflects the geometry: a back-door path enters T from behind — it begins with an arrow pointing into T — and represents a confounding route. By contrast, directed causal paths $T \to \cdots \to Y$ leave T through the front and carry the genuine effect.
Condition 2 is equivalent to requiring $(Y \perp\!\!\!\perp T \mid \mathbf{S})_{\mathcal{G}_{\underline{T}}}$: $\mathbf{S}$ d-separates T and Y in the graph obtained by deleting all arrows *out of* T (which eliminates causal paths, leaving only back-door paths). The graph $\mathcal{G}_{\overline{T}}$, which deletes arrows into T, is the wrong graph for this check.
Condition 1 adds an important constraint absent from Chapter 2: $\mathbf{S}$ must contain no descendant of T. Every valid back-door set is a d-separator in $\mathcal{G}_{\underline{T}}$, but not vice versa.

Example: A Single Confounder

Consider the DAG $T \leftarrow C \to Y$ with $T \to Y$, where C is observed. Does $\mathbf{S} = \{C\}$ satisfy the back-door criterion?

1. C is not a descendant of T.
2. $\{C\}$ blocks $T \leftarrow C \to Y$ (fork at C).

The back-door formula gives: $P(y \mid \mathrm{do}(T{=}t)) = \sum_c P(y \mid t, c)\, P(c)$.

Example: Two Back-Door Paths — Both Must Be Blocked

Consider treatment T, outcome Y, observed C_1 and C_2, unobserved U, with edges $C_1 \to T$, $C_1 \to Y$, $U \to T$, $U \to C_2$, $C_2 \to Y$, $T \to Y$. There are exactly two back-door paths:

1. $T \leftarrow C_1 \to Y$ (direct confounding by C_1)
2. $T \leftarrow U \to C_2 \to Y$ (indirect confounding via U routing through proxy C_2)

$\mathbf{S} = \{C_1\}$ fails: blocks path 1 but leaves path 2 open. $\mathbf{S} = \{C_2\}$ fails: blocks path 2 but leaves path 1 open. $\mathbf{S} = \{C_1, C_2\}$ succeeds:

$$P(y \mid \mathrm{do}(T{=}t)) = \sum_{c_1, c_2} P(y \mid t, c_1, c_2)\, P(c_1, c_2).$$

The unobserved U never appears: conditioning on C_2 blocks the path at the chain node, exploiting the *position* of the observed variables, not direct measurement of the confounders.

Why Condition 1 Is Essential: Conditioning on a Mediator

Consider a drug T that affects recovery Y via two routes: direct ($T \to Y$) and indirect through blood pressure M ($T \to M \to Y$), with unobserved confounder $U \to T$, $U \to Y$.
Naively trying $\mathbf{S} = \{M\}$ fails for two reasons: (1) M is a descendant of T, violating Condition 1; (2) conditioning on M blocks the causal pathway $T \to M \to Y$ while leaving the back-door path $T \leftarrow U \to Y$ untouched. The adjusted quantity $\sum_m P(y \mid t, m)P(m)$ is a causally ambiguous mixture of attenuated causal signal and uncontrolled confounding — neither the total causal effect nor the direct effect.

M-Bias: Conditioning on a Collider Opens a Closed Back-Door Path

Consider $T \to Y$ with $U_1 \to T$, $U_1 \to C$, $U_2 \to C$, $U_2 \to Y$ (U_1, U_2 unobserved, C observed). The only back-door path is $T \leftarrow U_1 \to C \leftarrow U_2 \to Y$, which has a collider at C. With $\mathbf{S} = \emptyset$, the path is blocked — the causal effect is already identified: $P(y \mid \mathrm{do}(t)) = P(y \mid t)$.
Including C in the adjustment set *activates* the collider, opening the previously dormant path. The set $\{C\}$ fails Condition 2: it does not block the back-door path; it creates one. This is called *M-bias* (from the M-shaped skeleton). The practical lesson: Condition 1 alone ("include all pretreatment variables") is not sufficient.

Theorem: Back-Door Adjustment Formula [@pearl1993comment]

If $\mathbf{S}$ satisfies the back-door criterion for the effect of T on Y, and $P(T{=}t \mid \mathbf{S}{=}\mathbf{s}) > 0$ for all $\mathbf{s}$ with $P(\mathbf{S}{=}\mathbf{s}) > 0$ (positivity), then:

$$P(y \mid \mathrm{do}(T{=}t)) = \int P(y \mid T{=}t,\ \mathbf{S}{=}\mathbf{s})\, dP(\mathbf{s}). \tag{3.1}$$

Equivalently: $\mathbb{E}[h(Y) \mid \mathrm{do}(T{=}t)] = \mathbb{E}_{\mathbf{S}}[\mathbb{E}[h(Y) \mid T{=}t,\ \mathbf{S}]]$.

Remark: Positivity Is a Data-Support Condition

The positivity condition requires every value of T to be observable within each stratum of $\mathbf{S}$. It is a property of the joint distribution P, not of the graph. Both the back-door criterion (graphical) and positivity (distributional) must hold simultaneously. When positivity fails on a set of measure zero, identification still holds in theory but sparse or empty cells pose practical difficulties.

3.2.1 The Adjustment Formula as Standardization

Equation 5.5 is also known as the *standardization formula* or *g-formula* (Robins 1986). The structure is important:

- $P(y \mid t, \mathbf{s})$ is the stratum-specific conditional distribution. Within each stratum, all back-door paths are blocked by $\mathbf{S}$, so this conditional distribution is causal.
- The outer integration $\int \cdots dP(\mathbf{s})$ weights strata by the *population marginal* distribution of $\mathbf{S}$, not the treatment-conditional $dP(\mathbf{s} \mid t)$. In a randomized experiment, $\mathbf{S} = \emptyset$ and Equation 5.5 reduces to $P(y \mid \text{do}(t)) = P(y \mid t)$.

Example: Back-Door Adjustment — A Numerical Walkthrough

Setup. Single confounder graph ($T \leftarrow C \rightarrow Y$, $T \rightarrow Y$), binary variables. C = disease severity (0 = mild, 1 = severe); T = treatment; Y = recovery. $P(C{=}1) = 0.5$, $P(T{=}1 \mid C{=}0) = 0.20$, $P(T{=}1 \mid C{=}1) = 0.80$.
Outcome probabilities. $P(Y{=}1 \mid T{=}1, C{=}0) = 0.55$, $P(Y{=}1 \mid T{=}0, C{=}0) = 0.45$, $P(Y{=}1 \mid T{=}1, C{=}1) = 0.45$, $P(Y{=}1 \mid T{=}0, C{=}1) = 0.35$.
Naïve (unadjusted) estimates. $P(Y{=}1 \mid T{=}1) = 0.50$, $P(Y{=}1 \mid T{=}0) = 0.39$, difference $= +0.11$.
Back-door adjusted estimates.

$$P(Y{=}1 \mid \text{do}(T{=}1)) = 0.55 \times 0.5 + 0.45 \times 0.5 = 0.50.$$

$$P(Y{=}1 \mid \text{do}(T{=}0)) = 0.45 \times 0.5 + 0.35 \times 0.5 = 0.40.$$

Causal risk difference $= 0.50 - 0.40 = 0.10$. The naïve estimate ($+0.11$) is close here because the confounding is mild; in general it can differ substantially.

3.3 The Front-Door Criterion

The back-door criterion requires observed variables to block every confounding path. When the confounder U is unobserved and no observed variable lies on the back-door path, a different strategy is needed: the *front-door criterion* intercepts the causal pathway at an observed mediator.

Definition: Front-Door Criterion

A set M of observed variables satisfies the **front-door criterion** for the effect of T on Y in $\mathcal{G}$ if:

1. *M intercepts all directed paths* from T to Y: every directed path from T to Y passes through some node in M.
2. There are *no unblocked back-door paths from T to M*: the path set $\{T \leftarrow \cdots \rightarrow M\}$ is empty or blocked.
3. *All back-door paths from M to Y are blocked by T*: conditioning on T blocks every path that begins with an arrow into M and ends at Y.

The front-door strategy avoids unobserved confounding by splitting the total effect into two identifiable pieces: (1) the effect of T on M (unconfounded by Condition 2), and (2) the effect of M on Y (made identifiable after conditioning on T by Condition 3).

Example: Smoking, Tar, and Cancer [@pearl1995causal]

T = smoking, M = tar deposits, Y = lung cancer, U = genetic predisposition (unobserved). Graph: $T \rightarrow M \rightarrow Y$ with $U \rightarrow T$ and $U \rightarrow Y$, no direct $T \rightarrow Y$ edge.
Conditions 1–3 hold: (1) the only directed path $T \rightarrow M \rightarrow Y$ passes through M; (2) no back-door path from T to M exists (no $U \rightarrow M$ edge); (3) the only back-door path from M to Y is $M \leftarrow T \leftarrow U \rightarrow Y$, which is blocked by conditioning on T.
The front-door formula applies:

$$P(y \mid \text{do}(T{=}t)) = \sum_{m} P(m \mid t) \sum_{t'} P(y \mid t', m)\, P(t').$$

Example: The Prototypical Front-Door Example

The simplest front-door graph is $T \to M \to Y$ with $U \to T$, $U \to Y$ (no direct $T \to Y$ edge). Checking the criterion:

1. The only directed path is $T \to M \to Y$, which passes through M.
2. No back-door path from T to M exists (no $U \to M$ edge).
3. The only back-door path from M to Y is $M \leftarrow T \leftarrow U \to Y$; conditioning on T blocks it.

Example: When the Front-Door Criterion Fails

Case (a): Condition 2 fails — the $T \to M$ link is confounded. Adding the edge $U \to M$: now the path $T \leftarrow U \to M$ is an unblocked back-door path from T to M. Stage 1 uses $P(M{=}m \mid T{=}t)$ as if the $T \to M$ link were unconfounded, but $U \to M$ means the association mixes cause and confounding.
Case (b): Condition 3 fails — the $M \to Y$ link has an extra confounder. Adding $V \to M$ and $V \to Y$ (V unobserved): the path $M \leftarrow V \to Y$ is not blocked by conditioning on T. Stage 2 can no longer use T as a valid back-door adjustment for $M \to Y$.

Modification	Condition violated	Stage broken
Add $U \to M$	Cond. 2: $T \to M$ confounded	Stage 1
Add $V \to M$, $V \to Y$	Cond. 3: $M \to Y$ confounded by V	Stage 2

In both cases the criterion detects the failure before any formula is written down.

3.3.1 The Front-Door Formula

Theorem: Front-Door Adjustment Formula [@pearl1995causal]

If M satisfies the front-door criterion for the effect of T on Y, and $P(t) > 0$ for all t, then:

$$P(y \mid \mathrm{do}(T{=}t)) = \sum_m \underbrace{P(m \mid T{=}t)}_{\text{Stage 1}} \underbrace{\sum_{t'} P(y \mid T{=}t',\, M{=}m)\, P(t')}_{\text{Stage 2}}. \tag{3.2}$$

Equivalently, writing $\mu(t', m) = \mathbb{E}[Y \mid T{=}t', M{=}m]$:

$$\mathbb{E}[Y \mid \mathrm{do}(T{=}t)] = \mathbb{E}_{M|T=t}[\mathbb{E}_T[\mu(T,\, M)]].$$

Reading the formula. Stage 1: $P(m \mid T{=}t)$ is the distribution of M given $T{=}t$. Condition 2 guarantees no confounding between T and M, so this is directly identifiable. Stage 2: $\sum_{t'} P(y \mid T{=}t', M{=}m)\, P(t')$ is the back-door-adjusted effect of M on Y, with T as the adjustment variable. Condition 3 guarantees that conditioning on T blocks all back-door paths from M to Y.

Heuristic derivation. The formula is the composition of two back-door applications. The role of each condition: Condition 1 together with Condition 2 allow decomposing through M; Condition 2 licenses replacing $\mathrm{do}(T{=}t)$ with conditioning in the M-marginal; Condition 3 licenses back-door adjustment for $M \to Y$.

3.4 The Three Rules of Do-Calculus

Both the back-door and front-door formulas are derivable from a single algebraic engine: the three rules of the do-calculus. Each rule is a conditional independence statement in a specific intervention graph.

Graph subscripts as bookkeeping. Each rule checks d-separation in a graph formed by specific edge deletions:

Subscript	Appears in	Arrows deleted
$\mathcal{G}_{\overline{X}}$	Rule 1	all arrows *into* X
$\mathcal{G}_{\overline{X}\underline{Z}}$	Rule 2	all into X; all out of Z
$\mathcal{G}_{\overline{X}\,\overline{Z(W)}}$	Rule 3	all into X; into Z-nodes not ancestors of W in $\mathcal{G}_{\overline{X}}$

When $W = \emptyset$, Rule 3 uses $\mathcal{G}_{\overline{X}\,\overline{Z}}$ (into both X and Z deleted). D-separation in the modified graph is checked by the same algorithm as Chapter 2.

Theorem: The Three Rules of Do-Calculus [@pearl1995causal]

Let $\mathcal{G}$ be a DAG over $\mathcal{V}$, and let $X, Y, Z, W \subseteq \mathcal{V}$ be disjoint. The rules hold for any distribution P Markov with respect to $\mathcal{G}$.

Rule 1 (Insertion/Deletion of Observations). *Remove or insert an observation when it becomes irrelevant in the post-intervention graph.*

$$P(y \mid \mathrm{do}(x), z, w) = P(y \mid \mathrm{do}(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z \mid X, W)_{\mathcal{G}_{\overline{X}}}. \tag{3.3}$$

Rule 2 (Action/Observation Exchange). *Replace an intervention by an observation, or vice versa, when the modified graph makes them equivalent.*

$$P(y \mid \mathrm{do}(x), \mathrm{do}(z), w) = P(y \mid \mathrm{do}(x), z, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z \mid X, W)_{\mathcal{G}_{\overline{X}\underline{Z}}}. \tag{3.4}$$

Rule 3 (Insertion/Deletion of Actions). *Remove or insert an intervention when it becomes irrelevant in the modified graph.*

$$P(y \mid \mathrm{do}(x), \mathrm{do}(z), w) = P(y \mid \mathrm{do}(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z \mid X, W)_{\mathcal{G}_{\overline{X}\,\overline{Z(W)}}}. \tag{3.5}$$

Remark: Soundness of the Rules

Each rule is *sound*: it holds for every distribution Markov with respect to $\mathcal{G}$. Rule 2's intuition is instructive: deleting Z's outgoing arrows in $\mathcal{G}_{\overline{X}\underline{Z}}$ simulates what intervening on Z would remove from Y's perspective. If Y is d-separated from Z after this deletion, then there is no causal channel whose behavior depends on whether Z was intervened on or merely observed.

3.4.1 Intuition for Each Rule

Rule 1 — Adding/Removing Observations. If Y and Z are independent given W in the post-do(x) world, then Z carries no additional information about Y and can be added or dropped from the conditioning set.

Rule 2 — Swapping Action for Observation. The most frequently used rule. Deleting Z's outgoing arrows cuts Z's causal paths to its descendants. If Y and Z are d-separated after this deletion, then intervention and observation on Z produce the same distribution of Y — because the only paths from Z to Y were Z's causal paths, which both operations sever (intervention) or leave intact (observation) equally.

Rule 3 — Deleting an Intervention. After deleting arrows into X and into the relevant Z-nodes, Z has no remaining path to Y. Setting Z to any value has no effect on Y, so do(z) can be dropped.

Example: Rule 1 — Deleting a Redundant Observation

Graph: $Z \to T \to Y$, no other edges. Intervene on T. In $\mathcal{G}_{\overline{T}}$: delete $Z \to T$; node Z is disconnected from Y. Hence $(Y \perp\!\!\!\perp Z)_{\mathcal{G}_{\overline{T}}}$. Rule 1 gives $P(y \mid \mathrm{do}(t), z) = P(y \mid \mathrm{do}(t))$: once T is fixed externally, its former cause Z carries no information about Y.

Example: Rule 2 — Action/Observation Exchange

Graph: $X \to Z \to Y$, no other edges. Form $\mathcal{G}_{\overline{X}\underline{Z}}$ (delete arrows into X — none — and arrows out of Z — removes $Z \to Y$). Node Z has no outgoing arrows; no path from Z to Y exists. Hence $(Y \perp\!\!\!\perp Z)_{\mathcal{G}_{\overline{X}\underline{Z}}}$. Rule 2 gives $P(y \mid \text{do}(x), \text{do}(z)) = P(y \mid \text{do}(x), z)$.

Example: Rule 3 — Deleting an Irrelevant Intervention

Graph: $Z \to X \to Y$, no other edges. Form $\mathcal{G}_{\overline{X}\,\overline{Z}}$ (delete into X — removes $Z \to X$ — and into Z — none). Node Z is isolated. Hence $(Y \perp\!\!\!\perp Z)_{\mathcal{G}_{\overline{X}\,\overline{Z}}}$. Rule 3 gives $P(y \mid \text{do}(x), \text{do}(z)) = P(y \mid \text{do}(x))$: once X is fixed, Z's only route to Y is severed.

Example: A Two-Rule Derivation — The Proxy Variable Graph

Graph. $T \to Y$, $U \to T$ (unobserved), $U \to S$ (observed), $S \to Y$.
Goal. Simplify $P(y \mid \text{do}(t))$ to an observational expression. S lies on the back-door path $T \leftarrow U \to S \to Y$ and is observed.
Step 1. Introduce S by total probability: $P(y \mid \text{do}(t)) = \sum_s P(y \mid \text{do}(t), s)\, P(s \mid \text{do}(t))$.
Step 2. Simplify $P(s \mid \text{do}(t))$ by Rule 3. The required graph is $\mathcal{G}_{\overline{T}}$. Because S is not a descendant of T, no path from T to S exists in $\mathcal{G}_{\overline{T}}$. Hence $(S \perp\!\!\!\perp T)_{\mathcal{G}_{\overline{T}}}$, and Rule 3 gives $P(s \mid \text{do}(t)) = P(s)$.
Step 3. Simplify $P(y \mid \text{do}(t), s)$ by Rule 2. The required graph is $\mathcal{G}_{\underline{T}}$. In $\mathcal{G}_{\underline{T}}$, the only paths between T and Y are back-door paths. The set $\{S\}$ d-separates T and Y in $\mathcal{G}_{\underline{T}}$ (blocking $T \leftarrow U \to S \to Y$ at S). Rule 2 gives $P(y \mid \text{do}(t), s) = P(y \mid t, s)$.
Combining: $P(y \mid \text{do}(t)) = \sum_s P(y \mid t, s)\, P(s)$. □

3.5 Do-Calculus Proofs of the Main Theorems

Proof of the Back-Door Formula via Do-Calculus

Graph: $T \to Y$, $\mathbf{S} \to T$, $\mathbf{S} \to Y$ (with $\mathbf{S}$ satisfying the back-door criterion).
Step 1. Introduce $\mathbf{S}$: $P(y \mid \text{do}(t)) = \int P(y \mid \text{do}(t), \mathbf{s})\, dP(\mathbf{s} \mid \text{do}(t))$.
Step 2. Rule 3 with $X = \emptyset$, $Z = T$, $W = \emptyset$: required graph $\mathcal{G}_{\overline{T}}$. Since $\mathbf{S}$ contains no descendants of T (back-door condition 1), $(\mathbf{S} \perp\!\!\!\perp T)_{\mathcal{G}_{\overline{T}}}$, so $P(\mathbf{s} \mid \text{do}(t)) = P(\mathbf{s})$.
Step 3. Rule 2 with $X = \emptyset$, $Z = T$, $W = \mathbf{S}$: required graph $\mathcal{G}_{\underline{T}}$. Back-door condition 2 states $(Y \perp\!\!\!\perp T \mid \mathbf{S})_{\mathcal{G}_{\underline{T}}}$, so $P(y \mid \text{do}(t), \mathbf{s}) = P(y \mid t, \mathbf{s})$.
Combining: $P(y \mid \text{do}(t)) = \int P(y \mid t, \mathbf{s})\, dP(\mathbf{s})$. □

Proof of the Front-Door Formula via Do-Calculus

Graph: $T \to M \to Y$, $U \to T$ (unobserved), $U \to Y$.
Step 1. Introduce M: $P(y \mid \text{do}(t)) = \int P(y \mid \text{do}(t), m)\, dP(m \mid \text{do}(t))$.
Step 2. Rule 2 to simplify $P(m \mid \text{do}(t))$: required graph $\mathcal{G}_{\underline{T}}$ (delete $T \to M$). In this graph no path from T to M remains (front-door condition 2 ensures no $U \to M$ edge), so $(M \perp\!\!\!\perp T)_{\mathcal{G}_{\underline{T}}}$. Rule 2 gives $P(m \mid \text{do}(t)) = P(m \mid t)$.
Step 3a. Convert conditioning on m to intervention. Rule 2 with $X = T$, $Z = M$, $W = \emptyset$: form $\mathcal{G}_{\overline{T}\,\underline{M}}$ (delete $U \to T$ and $M \to Y$). In this graph no path from M to Y exists, so $(Y \perp\!\!\!\perp M \mid T)_{\mathcal{G}_{\overline{T}\,\underline{M}}}$. Rule 2 gives $P(y \mid \text{do}(t), m) = P(y \mid \text{do}(t), \text{do}(m))$.
Step 3b. Drop the redundant $\text{do}(t)$. Rule 3 with $X = M$, $Z = T$, $W = \emptyset$: form $\mathcal{G}_{\overline{M}\,\overline{T}}$ (delete $T \to M$ and $U \to T$). Node T is isolated, so $(Y \perp\!\!\!\perp T)_{\mathcal{G}_{\overline{M}\,\overline{T}}}$. Rule 3 gives $P(y \mid \text{do}(t), \text{do}(m)) = P(y \mid \text{do}(m))$.
Step 3c. Back-door formula for $M \to Y$. By front-door condition 3, $\{T\}$ satisfies the back-door criterion for $M \to Y$: $P(y \mid \text{do}(m)) = \int P(y \mid t', m)\, dP(t')$.
Assembling: $P(y \mid \text{do}(t)) = \int [\int P(y \mid t', m)\, dP(t')] dP(m \mid t)$. □

3.6 The Do-Calculus Is Complete

A natural question: are the three rules enough? Could there be a graph where the causal effect is identifiable in principle, but no sequence of rules can derive it? Shpitser and Pearl (2006) answered this definitively.

Semi-Markovian DAGs encode latent common causes compactly as bidirected edges: $X \leftrightarrow Y$ stands for an unobserved U with $U \to X$ and $U \to Y$.

Theorem: Completeness of the Do-Calculus [@shpitser2006identification]

Let $\mathcal{G}$ be a semi-Markovian DAG. A causal quantity $P(y \mid \text{do}(t))$ is identifiable from P relative to $\mathcal{G}$ if and only if the do-calculus can derive a purely observational expression for it.

Example: The Bow Graph — The Simplest Non-Identifiable Structure

The *bow graph* has one directed edge $T \to Y$ and one bidirected edge $T \leftrightarrow Y$ (representing unobserved U with $U \to T$, $U \to Y$). No observed variable can block the back-door path $T \leftarrow U \to Y$; no observed mediator exists for front-door. We show directly that $P(y \mid \text{do}(t))$ is not identified by constructing two models that agree on $P(T, Y)$ but disagree on $P(Y \mid \text{do}(T{=}1))$.
Let $T, Y, U \in \{0, 1\}$ with $U \sim \text{Bernoulli}(1/2)$.
Model $\mathcal{M}_1$ (pure confounding; no causal effect): $T = U$, $Y = U$. Both T and Y are driven by U. The arrow $T \to Y$ carries no causal influence. Observed distribution: $P(T{=}0, Y{=}0) = P(T{=}1, Y{=}1) = 1/2$. Under $\text{do}(T{=}1)$: $P_1(Y{=}1 \mid \text{do}(T{=}1)) = P(U{=}1) = 1/2$.
Model $\mathcal{M}_2$ (full causal effect): $T = U$, $Y = T$. Y is determined entirely by T. Since $T = U$, the observed distribution is again $P(T{=}0, Y{=}0) = P(T{=}1, Y{=}1) = 1/2$ — *identical* to $\mathcal{M}_1$. Under $\text{do}(T{=}1)$: $P_2(Y{=}1 \mid \text{do}(T{=}1)) = 1$.
The two models produce identical observational distributions but assign different values (1/2 vs 1) to $P(Y{=}1 \mid \text{do}(T{=}1))$. No data set, however large, can distinguish them. The causal effect is *not identified*.

Proof Sketch and the ID Algorithm

Shpitser and Pearl (2006) introduced the *ID algorithm*: it takes a semi-Markovian DAG $\mathcal{G}$ and either returns an observational formula or reports non-identification.
Sufficiency (do-calculus identifies whenever possible): proved constructively. The key concept is the *c-component* — a maximal set of nodes connected by bidirected edges. The ID algorithm: (1) removes non-ancestors of Y; (2) decomposes over c-components of $\mathcal{G}_{\overline{T}}$; (3) recursively identifies each factor; (4) reports FAIL if any subproblem cannot be reduced.
Necessity (do-calculus cannot identify non-identifiable quantities): proved via the *hedge*. A hedge for $P(y \mid \text{do}(t))$ is a pair of subgraphs (F, F') encoding a "loop" of confounding the do-calculus cannot break. When a hedge exists, one can construct two models $\mathcal{M}_1$ and $\mathcal{M}_2$ with identical observed P but different $P(y \mid \text{do}(t))$. Combining: the do-calculus succeeds if and only if no hedge exists. □

Remark: Practical Implication of Completeness

If the ID algorithm reports non-identification, then *no estimation method — however clever — can recover the causal effect from observational data alone* given the assumed graph structure. Non-identification is not a limitation of a particular technique; it is a fundamental property of the causal model. Three remedies: impose additional structural assumptions (e.g., linearity, monotonicity); collect additional data (e.g., an instrument or a randomized experiment); or target a different, identified causal quantity (e.g., partial identification bounds, or the effect among the treated).

3.7 The Big Picture

This chapter completes the *theoretical* identification machinery of the course. The full causal inference pipeline is:

$$\text{SEM/DAG} \xrightarrow{\text{d-separation}} \text{conditional independences} \xrightarrow{\text{back-door, front-door, do-calculus}} P(y \mid \text{do}(t)) \xrightarrow{\text{estimation}} \widehat{\tau}.$$

The three identification criteria relate as follows. The *back-door criterion* is applicable when observed covariates can block all confounding paths; it is the workhorse of regression adjustment and propensity score methods (Chapters 5–6). The *front-door criterion* is applicable when an observed mediator intercepts all causal paths while remaining unconfounded from treatment; it is the mechanism behind the front-door formula (Chapter 8). The *do-calculus* subsumes both: it is complete in the sense of the Completeness Theorem, and its three rules constitute a universal derivation engine for identification.

3.8 Summary

Symbol	Meaning
$\mathcal{G}_{\overline{X}}$	Mutilated graph: all arrows into X deleted
$\mathcal{G}_{\underline{X}}$	Auxiliary graph: all arrows out of X deleted
Back-door criterion	Observed $\mathbf{S}$ not descending from T; blocks all back-door paths
Front-door criterion	Observed M intercepting all causal paths; satisfying three conditions
Rule 1	Delete/insert observations if d-separated in $\mathcal{G}_{\overline{X}}$
Rule 2	Swap action for observation if d-separated in $\mathcal{G}_{\overline{X}\,\underline{Z}}$
Rule 3	Delete action if d-separated in $\mathcal{G}_{\overline{X}\,\overline{Z(W)}}$

1. Graph surgery formalizes intervention: $\mathcal{G}_{\overline{T}}$ (delete arrows into T) represents do(T=t); d-separation in $\mathcal{G}_{\overline{T}}$ encodes conditional independence in $P(\cdot \mid \text{do}(t))$.
2. The back-door criterion identifies causal effects when observed variables block all confounding paths. Condition 1 (no descendants of T) and Condition 2 (d-separation in $\mathcal{G}_{\underline{T}}$) are both essential: the M-bias example shows that including a collider can introduce confounding.
3. The front-door criterion identifies causal effects via an observed mediator when direct confounding is unblockable. The two-stage structure of the formula corresponds to two unconfounded identification steps chained together.
4. The three rules of do-calculus — insertion/deletion of observations, action/observation exchange, and insertion/deletion of actions — are a sound and complete derivation engine for identification. The back-door and front-door formulas are special cases.
5. The completeness theorem of Shpitser and Pearl (2006) implies that non-identification is a property of the graph, not of the method: when the ID algorithm fails, no observational method can succeed.

3.9 Problems

1. Graph surgery and d-separation. Consider the DAG with edges $Z \to T$, $U \to T$, $U \to Y$, $T \to Y$ (U unobserved).

(a) Draw $\mathcal{G}_{\overline{T}}$ and $\mathcal{G}_{\underline{T}}$. For each graph, state which edges were deleted and why.
(b) In $\mathcal{G}_{\overline{T}}$, is $(Y \perp\!\!\!\perp Z)_{\mathcal{G}_{\overline{T}}}$? Justify by listing all paths and determining whether each is blocked.
(c) In the original $\mathcal{G}$, is $(Y \perp\!\!\!\perp Z \mid T)_{\mathcal{G}}$? How does this relate to the instruction "never condition on a mediator to remove confounding"?
(d) How does the absence of a direct edge $Z \to Y$ in the DAG relate to the exclusion restriction (the assumption that Z affects Y only through T)?

2. Back-door practice. Consider the DAG: $X \to T$, $X \to Y$, $T \to M$, $M \to Y$, $T \to Y$, where X is observed.

(a) List all back-door paths from T to Y.
(b) Does $\{X\}$ satisfy the back-door criterion for the effect of T on Y? Write the resulting adjustment formula.
(c) Does $\{M\}$ satisfy the back-door criterion? Explain why or why not.
(d) Does $\{X, M\}$ satisfy the back-door criterion? Identify which condition of the criterion M violates.
(e) Even if the formal criterion issue in (d) were set aside, explain the substantive bias that arises from conditioning on a mediator M that lies on a causal path $T \to M \to Y$.

3. Front-door identification. Consider the DAG: $T \to M \to Y$, with $U \to T$ and $U \to Y$ (unobserved U, no direct $T \to Y$ edge).

(a) Verify that M satisfies all three front-door conditions.
(b) Derive the front-door formula Equation 8.14 step by step, citing the do-calculus rule used at each step.
(c) Now add a direct edge $T \to Y$. Does M still satisfy the front-door criterion? Explain which condition fails.

4. Identification or non-identification? For each DAG, determine whether $P(y \mid \mathrm{do}(t))$ is identified non-parametrically. If identified, state the formula; if not, explain the obstruction.

(a) $T \to Y$, $U \to T$, $U \to Y$, U unobserved.
(b) Same as (a), with instrument $Z \to T$ and $Z \perp\!\!\!\perp U$ (no direct $Z \to Y$ edge). Show $P(y \mid \mathrm{do}(t))$ is *not* identified by constructing two binary SEMs $\mathcal{M}_1$ and $\mathcal{M}_2$, each compatible with the IV graph, such that $P_{\mathcal{M}_1}(Z, T, Y) = P_{\mathcal{M}_2}(Z, T, Y)$ but $P_{\mathcal{M}_1}(Y{=}1 \mid \mathrm{do}(T{=}1)) \neq P_{\mathcal{M}_2}(Y{=}1 \mid \mathrm{do}(T{=}1))$.
(c) $T \to M \to Y$, $U \to T$, $U \to Y$, $U \to M$, U unobserved.
(d) $T \to Y$, $X \to T$, $X \to Y$, X observed.

5. Back-door formula: proof and uniqueness.

(a) Under the conditions of **?@thm-backdoor**, prove Equation 5.5 using the three rules of do-calculus, following the proof sketch in Section 3.5.
(b) Suppose $\mathbf{S}_1$ and $\mathbf{S}_2$ both satisfy the back-door criterion (with positivity). Deduce from (a) that $\int P(y \mid t, \mathbf{s}_1)\, dP(\mathbf{s}_1) = \int P(y \mid t, \mathbf{s}_2)\, dP(\mathbf{s}_2)$. Interpret: the identification target $P(y \mid \mathrm{do}(t))$ is unique even when the adjustment set is not.

6. Rule sequencing on a graph with two observed confounders. Consider the DAG: $C_1 \to T$, $C_1 \to Y$, $U \to T$, $U \to C_2$, $C_2 \to Y$, $T \to Y$, with C_1, C_2 observed and U unobserved. Problem 2 established that $\{C_1, C_2\}$ is a valid back-door set. Derive $P(y \mid \mathrm{do}(t)) = \sum_{c_1, c_2} P(y \mid t, c_1, c_2)\, P(c_1, c_2)$ from first principles using the three rules. For each step: (i) name the rule applied, (ii) state which arrows are deleted and what edges remain, (iii) verify the required d-separation condition.

7. Non-identifiability by construction. Consider the graph from Problem 4(c): $T \to M \to Y$, $U \to T$, $U \to M$, $U \to Y$ (front-door condition 2 violated). Let $T, M, Y, U \in \{0, 1\}$ with $U \sim \mathrm{Bernoulli}(1/2)$.

(a) Construct two SEMs $\mathcal{M}_1$ and $\mathcal{M}_2$, each compatible with the graph, such that $P_{\mathcal{M}_1}(T, M, Y) = P_{\mathcal{M}_2}(T, M, Y)$ but $P_{\mathcal{M}_1}(Y{=}1 \mid \mathrm{do}(T{=}1)) \neq P_{\mathcal{M}_2}(Y{=}1 \mid \mathrm{do}(T{=}1))$.
(b) Explain which identification strategy fails: (i) back-door criterion, (ii) front-door criterion (which condition fails), (iii) completeness theorem (what does the existence of your two models imply?).

Part II

Identification

Chapter 4

Potential Outcomes and Adjustment

Learning Objectives

By the end of this chapter, students should be able to:

1. Define the potential outcome $Y(t)$, state SUTVA precisely, and derive the consistency equation $Y = Y(T)$.
2. Define the ATE and ATT, explain how they relate to each other, and re-express them as functionals of the structural equation for Y.
3. Show that the linear SEM structural coefficient β equals the ATE under homogeneous effects, and explain why ATE and ATT diverge under treatment effect heterogeneity.
4. State the ignorability assumption, interpret it graphically as a back-door condition, and explain why overlap is a necessary companion assumption.
5. Derive the adjustment formula for the ATE under strong ignorability and positivity, and explain how the back-door criterion provides the graphical justification for conditional exchangeability.

4.1 Motivation: A Third Language for Causality

The first three chapters developed causal inference primarily in the languages of structural equations and directed acyclic graphs. Those languages are especially effective for expressing interventions syntactically: the SEM shows how an intervention replaces an assignment mechanism, and the DAG shows how it deletes incoming arrows. This chapter introduces a third language: the *potential outcomes framework* of Neyman (1923) and Rubin (1974).

Potential outcomes provide a direct language for defining causal estimands. Quantities such as the average treatment effect, the average treatment effect on the treated, and related contrasts are most naturally written in terms of counterfactual outcomes $Y(1)$, $Y(0)$, and more generally $Y(t)$. For this reason, the potential outcomes framework has become standard in statistics, biostatistics, epidemiology, and much of econometrics.

Assumptions such as ignorability, positivity, and exclusion can be stated in this language, but DAGs and the do-operator often make their structural content more transparent. The two frameworks should be viewed as complementary rather than competing: *potential outcomes define the causal quantities of interest, while DAGs and the do-calculus clarify identification.*

4.2 The Neyman–Rubin Potential Outcomes Framework

4.2.1 The Potential Outcome

Definition: Potential Outcome [@neyman1923application; @rubin1974estimating]

For each unit i and each possible treatment value $t \in \mathcal{T}$, the *potential outcome* $Y_i(t)$ is the value of the outcome that unit i *would have exhibited* had its treatment been set to t, possibly contrary to fact.
The collection $\{Y_i(t) : t \in \mathcal{T}\}$ is called the *schedule of potential outcomes* for unit i. In the binary case $\mathcal{T} = \{0, 1\}$, the schedule is $(Y_i(0),\ Y_i(1))$.

$Y_i(1)$ is the outcome unit i would achieve under treatment; $Y_i(0)$ is the outcome under control. Only one of these is ever observed for any given unit. The unobserved potential outcome is called the *counterfactual*.

The Fundamental Problem of Causal Inference [@holland1986statistics]

For any unit i, at most one potential outcome $Y_i(t)$ is observed. The individual-level causal effect $Y_i(1) - Y_i(0)$ is therefore *never directly observable*. Causal inference is an inference problem precisely because this quantity must be recovered from a population of units rather than a single unit.

4.2.2 SUTVA and Consistency

Definition: SUTVA [@rubin1980randomization]

The *stable unit treatment value assumption* has two components:

1. **No interference.** The potential outcome $Y_i(t)$ depends only on unit i's own treatment: $Y_i(t_1, \dots, t_n) = Y_i(t_i)$.
2. **No hidden versions.** For each treatment level t, there is a single well-defined version. If two units both receive $T = 1$, the treatment is the same in both cases.

Under SUTVA, the observed outcome equals the potential outcome at the treatment actually received:

$$Y_i = Y_i(T_i). \tag{4.1}$$

This *consistency equation* is the bridge between the potential outcome world and the observed data world. It fails whenever SUTVA fails: if treatment spills over between units (e.g., vaccination provides herd immunity), or if the treatment label $T = 1$ covers multiple distinct interventions.

4.2.3 Connection to the Do-Operator

Proposition: Potential Outcomes and the Do-Operator

Under the structural causal model of Chapters 1–3, together with consistency and no interference (SUTVA):

$$Y(t) \overset{d}{=} Y \mid \mathrm{do}(T{=}t), \tag{4.2}$$

so that $P(Y(t) \le y) = P(Y \le y \mid \mathrm{do}(T{=}t))$ for every y. Equivalently, $\mathbb{E}[Y(t)] = \mathbb{E}[Y \mid \mathrm{do}(T{=}t)]$.

Proof sketch. In the mutilated SEM $\mathcal{G}_{\overline{T}}$, the structural equation for T is replaced by $T := t$ for every unit. The structural model then determines Y from t and unit i's own background variables — which is precisely what the potential outcomes framework defines as $Y_i(t)$. SUTVA's no-interference condition ensures $Y_i(t)$ does not depend on other units' treatment values, so the marginal distribution of Y in $\mathcal{G}_{\overline{T}}$ equals the marginal distribution of $Y(t)$. □

This equivalence lets us move freely between two notational traditions. When defining estimands such as $\mathbb{E}[Y(1) - Y(0)]$, potential-outcome notation is most natural. When proving identification results from a

graph, the do-operator is often more transparent: $P(y \mid \text{do}(t)) \neq P(y \mid T{=}t)$ whenever T is endogenous (established in Chapter 1), whereas $Y(t)$ carries no syntactic marker for this gap.

Remark

The SEM representation $Y_i(t) = g(t, \mathbf{X}_i, U_{Y,i})$ clarifies why the two frameworks are equivalent in content but different in emphasis. The do-calculus works with the interventional distribution $P(y \mid \text{do}(t))$ as a population-level object and asks when it can be recovered from observational data. The potential outcomes framework works with unit-level quantities $Y_i(t)$ and asks what population summaries (ATE, ATT) are scientifically meaningful. The SEM ties the two together.

4.3 Causal Estimands

4.3.1 The Average Treatment Effect and Its Relatives

Definition: ATE and ATT

For a binary treatment $T \in \{0, 1\}$:

- The *average treatment effect* (ATE): $\tau_{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)]$.
- The *average treatment effect on the treated* (ATT): $\tau_{\text{ATT}} = \mathbb{E}[Y(1) - Y(0) \mid T{=}1]$.

The ATE and ATT coincide only when the treatment effect does not depend on who selected into treatment — i.e., when $\mathbb{E}[Y(t) \mid T] = \mathbb{E}[Y(t)]$ for $t \in \{0, 1\}$. In a randomized experiment this holds by design.

Neither quantity is directly observable. The naïve estimator $\hat{\tau}_{\text{naive}} = \bar{Y}_{T=1} - \bar{Y}_{T=0}$ estimates $\mathbb{E}[Y \mid T{=}1] - \mathbb{E}[Y \mid T{=}0] \neq \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$, with the gap being the selection bias induced by endogeneity.

4.3.2 Causal Estimands as Functionals of the Structural Equation

In the SEM framework, the outcome is determined by a structural equation $Y = g(T, \mathbf{X}, U_Y)$, where U_Y collects all sources of variation not accounted for by $(T, \mathbf{X})$. The potential outcome under $\text{do}(T{=}t)$ is:

$$Y_i(t) = g(t, \mathbf{X}_i, U_{Y,i}). \tag{4.3}$$

Substituting into the definitions:

$$\tau_{\text{ATE}} = \mathbb{E}[g(1, \mathbf{X}, U_Y) - g(0, \mathbf{X}, U_Y)], \qquad \tau_{\text{ATT}} = \mathbb{E}[g(1, \mathbf{X}, U_Y) - g(0, \mathbf{X}, U_Y) \mid T{=}1]. \tag{4.4}$$

The linear SEM as a special case. In the Gaussian linear SEM $Y = \beta T + \gamma^\top \mathbf{X} + \varepsilon$, the unit-level effect is $Y_i(1) - Y_i(0) = \beta$ for every unit. Consequently, $\tau_{\text{ATE}} = \tau_{\text{ATT}} = \beta$. The structural coefficient β *is* the average treatment effect. This homogeneity is a special property of the linear additive model, not a general feature.

Heterogeneous effects. In the nonparametric SEM, the unit-level effect $g(1, \mathbf{X}_i, U_{Y,i}) - g(0, \mathbf{X}_i, U_{Y,i})$ varies across units. The ATE and ATT differ whenever treatment selection correlates with individual effect size — i.e., whenever units who benefit more also tend to self-select into treatment.

4.4 Ignorability, Positivity, and Adjustment

4.4.1 Strong Ignorability

The central identifying assumption in observational studies is that treatment assignment is as good as random after conditioning on observed covariates X.

Definition: Strong Ignorability [@rosenbaum1983central]

The treatment assignment T is *strongly ignorable* given X if:

1. **Unconfoundedness:** $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$.

2. **Overlap (positivity):** $0 < P(T{=}1 \mid X{=}x) < 1$ for all x in the support of X.

4.4.2 Adjustment Formula under Ignorability

Under strong ignorability, the ATE is identified by the standardization formula:

$$\tau_{\text{ATE}} = \mathbb{E}_X[\mathbb{E}[Y \mid T{=}1, X] - \mathbb{E}[Y \mid T{=}0, X]]\,. \tag{4.5}$$

Derivation. For each treatment level t, by the law of iterated expectations:

$$\mathbb{E}[Y(t)] = \mathbb{E}_X[\mathbb{E}[Y(t) \mid X]]\,.$$

Under unconfoundedness, $\mathbb{E}[Y(t) \mid X] = \mathbb{E}[Y(t) \mid T{=}t, X]$. By consistency, $\mathbb{E}[Y(t) \mid T{=}t, X] = \mathbb{E}[Y \mid T{=}t, X]$. Therefore $\mathbb{E}[Y(t)] = \mathbb{E}_X[\mathbb{E}[Y \mid T{=}t, X]]$. Applying once for $t = 1$ and once for $t = 0$ and taking the difference yields Equation 4.5. □

Example: A Two-Stratum Adjustment Calculation

Let $X \in \{0, 1\}$ with $P(X{=}1) = 0.4$, $P(X{=}0) = 0.6$, and observed conditional means:

	$T = 1$	$T = 0$
$X = 1$	$\mathbb{E}[Y \mid T = 1, X = 1] = 8$	$\mathbb{E}[Y \mid T = 0, X = 1] = 6$
$X = 0$	$\mathbb{E}[Y \mid T = 1, X = 0] = 5$	$\mathbb{E}[Y \mid T = 0, X = 0] = 4$

Under ignorability and positivity:

$$\mathbb{E}[Y(1)] = 0.4 \times 8 + 0.6 \times 5 = 6.2, \qquad \mathbb{E}[Y(0)] = 0.4 \times 6 + 0.6 \times 4 = 4.8.$$

Hence $\tau_{\text{ATE}} = 6.2 - 4.8 = 1.4$. The formula works by comparing treated and control outcomes within each stratum and averaging those within-stratum comparisons over the marginal distribution of X.

4.4.3 Back-Door Interpretation

The potential-outcomes assumption $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$ is the counterfactual expression of the idea that, after conditioning on X, treatment assignment carries no residual information about the outcome that would be observed under intervention $T{=}t$. In DAG language, the closely related condition is that X blocks all back-door paths from T to Y.

Proposition: Back-Door Criterion Implies Ignorability

Under the NPSEM/SWIG semantics adopted in these notes, if X satisfies the back-door criterion for the effect of T on Y — i.e., (1) no node in X is a descendant of T, and (2) X blocks every back-door path from T to Y — then unconfoundedness $(Y(t) \perp\!\!\!\perp T \mid X)$ holds for all t.

Proof sketch. The target statement $Y(t) \perp\!\!\!\perp T \mid X$ is *cross-world*: it mixes the counterfactual $Y(t)$ with the factual treatment T, and is not a d-separation statement in $\mathcal{G}$ itself. The rigorous translation proceeds through the single-world intervention graph $\mathcal{G}(t)$ (Appendix B). Under the NPSEM semantics, $Y(t)$ and T depend on disjoint exogenous errors together with their own ancestors, and the back-door criterion is precisely the d-separation condition on $\mathcal{G}(t)$ that makes these conditionally independent given X. See Appendix B for the explicit SWIG derivation. □

Remark

This is the direction most important for practice: a graphical adjustment set justifies the counterfactual independence needed for standardization, regression adjustment, and propensity-score methods.

The converse — whether every ignorability statement corresponds to a back-door condition — is more delicate; see Appendix B.

Example: Labor Training Program

Following LaLonde (1986) and Dehejia and Wahba (1999), let T = receipt of job training, Y = earnings two years later, X = (age, education, prior earnings, race, marital status).
The back-door path $T \leftarrow X \rightarrow Y$ is blocked by conditioning on X, leaving only the causal path $T \rightarrow Y$. Ignorability holds if this DAG is correctly specified. If there is an unobserved variable U (motivation, ability) that affects both training participation and earnings, X fails the back-door criterion and ignorability fails.

4.4.4 Overlap and Positivity

Why Overlap Is Non-Negotiable

Without overlap, some covariate strata contain only treated or only control units. In those strata, $\mathbb{E}[Y \mid T{=}0, X{=}x]$ or $\mathbb{E}[Y \mid T{=}1, X{=}x]$ is unobservable, and Equation 4.5 cannot be evaluated at those values of x. Identification of the ATE fails not because the causal structure is wrong, but because the data do not span the support needed.
The ATT can remain identified under a *weaker, one-sided* overlap condition: $P(T{=}0 \mid X{=}x) > 0$ almost surely on the support of X in the treated subpopulation. Full two-sided overlap is not required for ATT, because ATT averages only over the treated population.

4.5 Where the Frameworks Agree and Diverge

Task	Potential Outcomes	Do-Calculus / DAG	SEM
Define ATE, ATT	Most natural notation	Via $\mathbb{E}[Y \mid \text{do}(t)]$	Via structural equations
Encode causal assumptions	Typically informal; graph implicit	Explicit directed edges	Structural equations
Read conditional independence	Requires auxiliary graph	d-separation	With graph only
Identification from data	Via ignorability	Back-door, front-door, do-calculus	Via exclusion restrictions
Likelihood construction	Estimating equations / semiparametric	Observational functionals only	Structural model
Cross-world assumptions	Natural to state	See Appendix B	Implicit in model
Standard in statistics	Dominant	Growing rapidly	Econometrics

Our Position in This Course

Both frameworks are indispensable. We use *potential outcomes* to *define* causal estimands. We use the *do-calculus and DAGs* for *identification*. The back-door criterion is the graphical condition that justifies conditional ignorability and hence the adjustment formula. The do-operator language is preferred throughout this course because it makes the interventional/observational distinction *syntactically enforced*: $P(y \mid \text{do}(t))$ cannot be confused with $P(y \mid T{=}t)$, whereas $\mathbb{E}[Y(t)]$ carries no such syntactic marker.

4.6 Summary

Causal inference = counterfactual questions + graphical assumptions + statistical estimation.

1. The potential outcome $Y(t)$ is the outcome that would be observed under intervention $T = t$. Under consistency, no interference, and the structural causal semantics adopted in these notes, the observed outcome satisfies $Y_i = Y_i(T_i)$, and $Y(t)$ has the same distribution as $Y \mid \mathrm{do}(T{=}t)$.
2. The ATE and ATT are averages of the unit-level causal effect $g(1, \mathbf{X}_i, U_{Y,i}) - g(0, \mathbf{X}_i, U_{Y,i})$ over the full population and the treated subpopulation respectively. In the linear SEM, both equal β. They diverge under heterogeneous effects when treatment selection correlates with individual effect size.
3. Strong ignorability $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$ with overlap identifies the ATE via Equation 4.5. The back-door criterion provides a sufficient graphical condition for the conditional exchangeability assumption.
4. Single world intervention graphs (Richardson and Robins 2014) provide a formal graphical representation of counterfactual variables, making ignorability a d-separation statement (Appendix B).
5. The frameworks are complementary: potential outcomes define estimands; the do-calculus identifies them. The back-door criterion connects the two by providing the graphical condition under which adjustment recovers the causal effect.

From ignorability to randomization. In observational studies, ignorability must be justified by substantive knowledge encoded in a causal graph. Randomized experiments provide a different solution: when treatment is assigned randomly, $(Y(0), Y(1)) \perp\!\!\!\perp T$ holds by design, guaranteeing ignorability *without any covariate adjustment*. Chapter 5 studies this design.

Identification vs. estimation. Chapters 1–4 have focused on *identification*: whether $\tau = \mathbb{E}[Y(1) - Y(0)]$ can be written as a functional $\Phi(P(Y, T, X))$ of the observed distribution. Estimation — constructing $\hat{\tau}$ from a finite sample to approximate $\Phi(P)$ — is studied in Part III, covering regression adjustment, IPW, doubly robust estimators, and instrumental variables.

4.7 Problems

1. SUTVA and consistency. Suppose $n = 3$ units receive binary treatments (T_1, T_2, T_3) and unit i's outcome may depend on all three treatments: $Y_i = Y_i(T_1, T_2, T_3)$.

(a) How many potential outcomes does unit 1 have? Write them out explicitly.
(b) SUTVA imposes no interference: $Y_i(T_1, T_2, T_3) = Y_i(T_i)$. How many distinct potential outcomes remain?
(c) Give a real-world example where no-interference plausibly holds and one where it plausibly fails. In the latter case, explain what identification strategy (if any) remains available.

2. ATE, ATT, and selection bias. Let $(Y(0), Y(1), T) \sim P$ with $P(T{=}1) = 0.5$, $\mathbb{E}[Y(1)] = 3$, $\mathbb{E}[Y(0)] = 1$, $\mathbb{E}[Y(1) \mid T{=}1] = 4$, $\mathbb{E}[Y(0) \mid T{=}1] = 2$, $\mathbb{E}[Y(1) \mid T{=}0] = 2$, $\mathbb{E}[Y(0) \mid T{=}0] = 0$.

(a) Compute the ATE and ATT. Are they equal?
(b) Compute the naïve estimator $\mathbb{E}[Y \mid T{=}1] - \mathbb{E}[Y \mid T{=}0]$. Decompose the gap between this and the ATE into a selection bias term and an ATT–ATE difference.
(c) Under what graphical condition on the DAG would the naïve estimator equal the ATE? State this condition in both potential outcome language (ignorability) and do-calculus language (back-door criterion).

3. Causal estimands from the structural equation. Consider the nonparametric SEM $Y = g(T, X, U_Y)$ with binary $T \in \{0, 1\}$, observed covariate X, and U_Y independent of (T, X) in the mutilated graph.

(a) Write τ_{ATE} and τ_{ATT} as expectations of $g(1, X, U_Y) - g(0, X, U_Y)$ over the appropriate distribution. Under what condition do they coincide?
(b) Specialize to the linear SEM $g(t, x, u) = \alpha + \beta t + \gamma x + u$. Show that the unit-level causal effect $Y_i(1) - Y_i(0)$ is constant across all units, and hence $\tau_{\mathrm{ATE}} = \tau_{\mathrm{ATT}} = \beta$.
(c) Now consider the heterogeneous SEM $g(t, x, u) = (\alpha + u)\, t + \gamma x$, where $U_Y \sim \mathcal{N}(0, \sigma^2)$, $\mathrm{Cov}(U_Y, T) = \rho$, and $P(T{=}1) = p \in (0, 1)$. Compute τ_{ATE} and τ_{ATT}. Show that they differ when $\rho \neq 0$, and interpret this difference.
(d) In part (c), show that the population OLS slope on T in the regression of Y on $(1, T, X)$ equals $\alpha + \rho/p = \tau_{\mathrm{ATT}}$, not $\tau_{\mathrm{ATE}} = \alpha$. What additional structure would be needed to identify τ_{ATE}?

4. SWIGs and ignorability. *(Requires Appendix B.)* Consider the DAG: $X \to T$, $X \to Y$, $T \to Y$, with X fully observed.

(a) Construct the SWIG $\mathcal{G}(t)$ by splitting T into its random and fixed halves. Draw the result, labeling the random half, fixed half, and the potential outcome $Y(t)$.

(b) In $\mathcal{G}(t)$, identify all paths between the random half T and $Y(t)$. Determine which are blocked and which are open before conditioning.

(c) Use d-separation in $\mathcal{G}(t)$ to verify that $(Y(t) \perp\!\!\!\perp T \mid X)_{\mathcal{G}(t)}$ holds. Conclude that X satisfies the back-door criterion, confirming (**prop-ign-bd?**).

(d) Now add a hidden common cause $U \to T$, $U \to Y$. Draw the revised SWIG. Does the ignorability argument still hold? State the correct conclusion and identify what additional structure would be needed for identification.

Chapter 5

Randomization and Back-Door Adjustment

Learning Objectives

By the end of this chapter, students should be able to:

1. Explain in do-calculus terms why a randomized experiment makes $f(y \mid \text{do}(T{=}t)) = f(y \mid T{=}t)$, and identify the graphical feature that produces this equality.
2. State the ignorability assumption in both potential-outcomes and graphical language, and distinguish weak from strong ignorability.
3. Apply the regression-adjusted estimator to reduce variance in a randomized experiment, and explain why consistency holds even under model misspecification.
4. Derive and compute the standardization (g-formula) estimator of the ATE from a contingency table or regression output.
5. Explain the method of stratification (subclassification) and describe when and why it removes confounding bias.
6. Identify Simpson's paradox in a numerical example and resolve it using the back-door criterion.
7. Articulate why propensity score methods are needed when X is high-dimensional, motivating Chapter 6.

5.1 Randomized Experiments

5.1.1 The Do-Calculus of Randomization

The defining feature of a randomized experiment is that the analyst *sets* the treatment for each unit, independently of all background variables. In do-calculus terms, the data come from the mutilated distribution $f(y \mid \text{do}(T{=}t))$ rather than the observational distribution $f(y \mid T{=}t)$.

The Fundamental Equality of Randomized Experiments

In a completely randomized experiment, T is assigned independently of all pre-treatment variables. In the DAG, this means there are *no arrows into* T: the treatment node has no parents. By Rule 2 of the do-calculus, with $X = \emptyset$, $Z = T$, $W = \emptyset$, the graphical condition $(Y \perp\!\!\!\perp T)_{\mathcal{G}_{\underline{T}}}$ holds because T is isolated in $\mathcal{G}_{\underline{T}}$ (no parents in $\mathcal{G}$, and the outgoing edge $T \to Y$ is deleted). Rule 2 therefore gives:

$$f(y \mid \text{do}(T{=}t)) = f(y \mid T{=}t). \tag{5.1}$$

In a randomized experiment, observing $T = t$ is the same as intervening to set $T = t$.

What the equality assumes beyond the graph. The fundamental equality Equation 5.1 relies on more than deletion of arrows into T. Two further conditions are needed. First, SUTVA (no interference

and no hidden treatment versions), which licenses the consistency equation $Y_i = Y_i(T_i)$. Second, the realized treatment must equal the assigned treatment for every unit (full compliance); when this fails, $f(y \mid T{=}t)$ describes outcomes among those who *actually* received t rather than those who were *assigned* t. When compliance fails, the intent-to-treat versus per-protocol distinction becomes substantive (Chapters 7 and 13).

Graphical argument. In the observational DAG, arrows into T from observed covariates X and unobserved confounders U create back-door paths from T to Y. Randomization *physically* severs these arrows: assignment is determined by a coin flip, not by X or U. The mutilated graph $\mathcal{G}_{\overline{T}}$ is, in a randomized experiment, the *actual* data-generating graph. There are no back-door paths to block, because none exist.

Potential outcomes statement. In the potential outcomes language, randomization implies $(Y(0), Y(1)) \perp\!\!\!\perp T$ unconditionally — no covariate adjustment is needed. This is the strongest possible version of ignorability.

5.1.2 Estimation of the ATE in a Randomized Experiment

Lemma: Identification under Complete Randomization

Under complete randomization, $\mathbb{E}[Y(t)] = \mathbb{E}[Y \mid T{=}t]$, $t \in \{0, 1\}$.

Proof. By the PO–do equivalence, $\mathbb{E}[Y(t)] = \mathbb{E}[Y \mid \mathrm{do}(T{=}t)]$. The treatment node has no parents, so Equation 5.1 gives $f(y \mid \mathrm{do}(T{=}t)) = f(y \mid T{=}t)$. □

Under complete randomization, the ATE is estimated by the *difference-in-means* (DIM) estimator:

$$\hat{\tau}_{\mathrm{DIM}} = \bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1} \sum_{i:\, T_i=1} Y_i - \frac{1}{n_0} \sum_{i:\, T_i=0} Y_i.$$

Theorem: Neyman's Theorem for the Difference-in-Means Estimator

Under complete randomization, let $\mathcal{F}_N = \{(Y_i(0), Y_i(1))\}$ be the potential outcomes (fixed), and let expectations be over the randomization distribution. Define the finite-population variances $S_t^2 = \frac{1}{n-1} \sum_i (Y_i(t) - \bar{Y}(t))^2$, cross-variance $S_{01} = \frac{1}{n-1} \sum_i (Y_i(0) - \bar{Y}(0))(Y_i(1) - \bar{Y}(1))$, and effect variance $S_\tau^2 = \frac{1}{n-1} \sum_i (\tau_i - \bar{\tau}_n)^2$. Then:

1. **Unbiasedness.** $\mathbb{E}[\hat{\tau}_{\mathrm{DIM}} \mid \mathcal{F}_N] = \bar{Y}(1) - \bar{Y}(0) = \bar{\tau}_n$.
2. **Design variance.**
$$\mathrm{Var}(\hat{\tau}_{\mathrm{DIM}} \mid \mathcal{F}_N) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\tau^2}{n}. \tag{5.2}$$
3. **Conservative variance estimator.** The within-arm sample variance estimator $\hat{V} = \hat{S}_1^2/n_1 + \hat{S}_0^2/n_0$ satisfies $\mathbb{E}[\hat{V} \mid \mathcal{F}_N] - \mathrm{Var}(\hat{\tau}_{\mathrm{DIM}} \mid \mathcal{F}_N) = S_\tau^2/n \geq 0$. It is exact when the unit-level treatment effects τ_i are constant.

Proof of Neyman's Theorem

Throughout we condition on $\mathcal{F}_N$. Under CRE, $\mathbb{E}[T_i \mid \mathcal{F}_N] = n_1/n$ for all i, and:

$$\mathrm{Cov}(T_i, T_j \mid \mathcal{F}_N) = \begin{cases} f_1(1 - f_1) & i = j, \\ -f_1(1 - f_1)/(n-1) & i \neq j, \end{cases}$$

where $f_1 = n_1/n$. The $i \neq j$ expression follows from sampling n_1 units without replacement.
(i) Unbiasedness. $\mathbb{E}[n_1^{-1} \sum_i T_i Y_i(1) \mid \mathcal{F}_N] = \bar{Y}(1)$ by $\mathbb{E}[T_i \mid \mathcal{F}_N] = n_1/n$; the control arm is symmetric.

(ii) Variance. Writing $Z_i(t) = Y_i(t) - \bar{Y}(t)$ and using the covariance formula:

$$\mathrm{Var}\left(\frac{1}{n_1}\sum_i T_i Y_i(1)\right) = \frac{f_1(1-f_1)}{n_1^2} \cdot \frac{n}{n-1}\sum_i Z_i(1)^2 = \frac{n_0}{n_1 n} S_1^2,$$

and the cross-arm covariance yields $-S_{01}/n$. Combining gives the first form of Equation 5.2. Rearranging using $n_0/(n_1 n) = 1/n_1 - 1/n$ gives the second form.
(iii) Conservatism. By finite-population sampling theory, $\mathbb{E}[\hat{S}_t^2 \mid \mathcal{F}_N] = S_t^2$. Therefore $\mathbb{E}[\hat{V} \mid \mathcal{F}_N] = S_1^2/n_1 + S_0^2/n_0$, and subtracting Equation 5.2 gives $S_\tau^2/n \geq 0$. □

Remark: Why the Conservatism Cannot Be Removed

The correction term S_τ^2/n depends on the unit-level effect variance. Computing S_τ^2 requires both $Y_i(0)$ and $Y_i(1)$ for every unit — precisely what the fundamental problem of causal inference forbids. The estimator $\hat{V}$ is not merely pragmatic; it is the sharpest variance estimator based on observed data alone without further modelling assumptions.

Remark: Superpopulation Variance as a Corollary

Under i.i.d. superpopulation draws with arm variances (σ_0^2, σ_1^2), the total variance has two phases: randomization variance (given $\mathcal{F}_N$) and sampling variance (over draws of $\mathcal{F}_N$). The law of total variance gives:

$$\mathrm{Var}(\hat{\tau}_{\mathrm{DIM}}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}. \tag{5.3}$$

The Neyman correction S_τ^2/n in Equation 5.2 is exactly cancelled by the sampling variance of $\bar{\tau}_n$ when averaging over superpopulation draws. See Ding (2024) for a complete treatment.

5.1.3 Fisher's Randomization Inference vs. Neyman's Repeated Sampling

Fisher's framework tests the *sharp null* $Y_i(1) = Y_i(0)$ for all i: under this null, every missing potential outcome is known, making the exact randomization distribution of any test statistic computable over all $\binom{n}{n_1}$ assignments. Neyman's framework studies the repeated-sampling behavior of estimators like $\hat{\tau}_{\mathrm{DIM}}$ for average causal effects.

Fisher's approach is aligned with exact hypothesis testing under a sharp null; Neyman's is aligned with point estimation and uncertainty quantification. Both rely on the treatment assignment mechanism but answer different questions.

Remark: Randomization Licenses Inference Without Population Assumptions

Randomization licenses causal inference *without any assumption about a population model.* In Fisher's design-based framework, the potential outcomes $\{Y_i(0), Y_i(1)\}$ are fixed numbers; the only source of randomness is the assignment vector $\mathbf{T}$, drawn uniformly from all $\binom{n}{n_1}$ possible assignments. Probability statements refer to this mechanism — not to repeated draws from a population. Units need not be a random sample; they can be a convenience sample, a census, or a targeted group. Observational methods, by contrast, must invoke either an i.i.d. sampling assumption or a superpopulation framework, because no analogous physical mechanism anchors probability statements.

Remark: Fisher's Sharp Null — Unidentifiable Effects, Exactly Testable Hypothesis

Individual causal effects $\tau_i = Y_i(1) - Y_i(0)$ are unidentifiable for any unit. Yet the sharp null $H_0 : Y_i(1) = Y_i(0)$ for all i is exactly testable. The resolution: under H_0, the sharp null completely specifies the joint potential outcome table — $Y_i(1) = Y_i(0) = Y_i^{\mathrm{obs}}$ regardless of assignment — so

the full $2n$-vector of potential outcomes is known, and the permutation distribution of any test statistic is computable from observed data with no estimation step.

Remark: Further Reading

Ding (2024) provides a comprehensive graduate-level treatment of CRE, stratified experiments, and regression adjustment. Li and Ding (2017) works out the asymptotic theory underlying confidence intervals for $\hat{\tau}_{\text{DIM}}$ via a finite-population CLT requiring neither i.i.d. observations nor a superpopulation model. Ding et al. (2016) develop randomization-based tests for treatment effect heterogeneity.

5.2 Ignorability

5.2.1 Two Routes to the Same Estimand

Section Section 5.1 established that randomization achieves Equation 5.1 and that DIM is unbiased for the ATE without covariate adjustment. In observational studies, no such physical severance occurs: treatment is selected based on characteristics that may include unmeasured variables U affecting the outcome.

Both settings target the same estimand — the ATE $\mathbb{E}[Y(1) - Y(0)]$ — but reach it by fundamentally different routes:

	Randomized experiment	**Observational study**
How ignorability arises	By design: researcher severs all arrows into T	By assumption: analyst asserts no unmeasured confounders
$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$	*Guaranteed* — holds unconditionally	*Assumed* — requires X to capture every confounder
Credibility	As strong as the randomization protocol	As strong as substantive knowledge of the DGP
Testability	Can audit the assignment mechanism	Cannot be verified from data alone
Failure mode	Protocol violations, non-compliance	Any unmeasured variable affecting both T and Y

Statistical adjustment can *implement* ignorability once assumed, but cannot *create* it. No covariate adjustment — however flexible — can close a back-door path through an unmeasured variable. This is why a well-conducted RCT is considered more credible than even the most carefully adjusted observational study.

5.2.2 Terminology

Strong ignorability was defined in Chapter 4: joint unconfoundedness $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$ together with overlap $0 < P(T{=}1 \mid X) < 1$ a.s. (Rosenbaum and Rubin 1983). The pointwise (weak) form $Y(t) \perp\!\!\!\perp T \mid X$ for each t separately is what identification of $\mathbb{E}[Y(t)]$ actually uses. Equivalent names in the literature: *unconfoundedness*, *selection on observables*, *no unmeasured confounders*, *conditional exchangeability*.

5.2.3 Three Languages for Ignorability

Language	**Statement of ignorability**
SEM	In $Y = f_Y(T, X, U_Y)$ with $T = f_T(X, U_T)$, the exogenous factor U_Y is independent of T given X — no unobserved common cause of T and Y remains after conditioning on X.

Language	Statement of ignorability
DAG / do-calculus	X satisfies the back-door criterion for (T, Y): X blocks all back-door paths and contains no descendant of T.
Potential outcomes	$Y(t) \perp\!\!\!\perp T \mid X$ for $t \in \{0, 1\}$ (weak ignorability).

These three forms are closely aligned under NPSEM-IE semantics together with consistency and no interference.

5.2.4 What Ignorability Requires

Ignorability Is Untestable from Observational Data

No statistical test can confirm or refute unconfoundedness using observed data alone. If an unobserved variable U affects both T and Y, the back-door path $T \leftarrow U \rightarrow Y$ is open, and conditioning on any set of observed variables cannot close it. Sensitivity analysis (Rosenbaum 2002) can quantify how large such an unobserved confounder would have to be to overturn a conclusion, but cannot confirm the assumption itself.

5.2.5 From Ignorability to Identification

Lemma: Identification of $\mathbb{E}[Y(t)]$

Suppose: (i) *weak ignorability* $Y(t) \perp\!\!\!\perp T \mid X$; (ii) *overlap* $P(T{=}t \mid X{=}x) > 0$ a.s.; (iii) *consistency* $Y = Y(T)$. Then:

$$\mathbb{E}[Y(t)] = \mathbb{E}\big[\mathbb{E}[Y \mid T{=}t, X]\big] = \int \mathbb{E}[Y \mid T{=}t, X{=}x]\, p(x)\, dx. \tag{5.4}$$

Proof. Apply the law of iterated expectations, then use each assumption in turn:

$$\mathbb{E}[Y(t)] \overset{\text{LIE}}{=} \mathbb{E}[\mathbb{E}[Y(t) \mid X]] \overset{(i)}{=} \mathbb{E}[\mathbb{E}[Y(t) \mid T{=}t, X]] \overset{(iii)}{=} \mathbb{E}[\mathbb{E}[Y \mid T{=}t, X]]. \quad \square$$

Overlap (ii) guarantees $\mathbb{E}[Y \mid T{=}t, X{=}x]$ is well-defined everywhere in the support of X.

Remark: Two Languages, One Formula

?@lem-ident and the back-door theorem (Chapter 3) are two proofs of the same formula in different languages. The graphical proof applies Rules 2 and 3 to the mutilated graph; the potential-outcomes proof applies the law of iterated expectations to counterfactual variables. The bridge is the structural causal semantics: under NPSEM-IE, $Y(t)$ has the same distribution as Y under $\text{do}(T{=}t)$ (Chapter 4). Consistency connects $\mathbb{E}[Y(t) \mid T{=}t, X]$ to the observed $\mathbb{E}[Y \mid T{=}t, X]$.

Theorem: Identification of the ATE and ATT

Under the same three assumptions:

$$\tau_{\text{ATE}} = \int \big[\mathbb{E}(Y \mid T{=}1, X{=}x) - \mathbb{E}(Y \mid T{=}0, X{=}x)\big]\, p(x)\, dx, \tag{5.5}$$

$$\tau_{\text{ATT}} = \int \big[\mathbb{E}(Y \mid T{=}1, X{=}x) - \mathbb{E}(Y \mid T{=}0, X{=}x)\big]\, p(x \mid T{=}1)\, dx. \tag{5.6}$$

Proof. Apply **?@lem-ident** separately to $t = 1$ and $t = 0$ and subtract. For the ATT, replace $p(x)$ by $p(x \mid T{=}1)$. $\square$

Assumption traceability. Each step in the proof invokes exactly one assumption:

Proof step	Assumption invoked	Failure mode
$\mathbb{E}[Y(t) \mid X] = \mathbb{E}[Y(t) \mid T{=}t, X]$	Weak ignorability	Unmeasured confounder: $Y(t)$ depends on T within X-strata
$\mathbb{E}[Y \mid T{=}t, X{=}x]$ is well-defined	Overlap	Empty stratum: no units with $T{=}t$ at x
$\mathbb{E}[Y(t) \mid T{=}t, X] = \mathbb{E}[Y \mid T{=}t, X]$	Consistency	Interference or hidden treatment versions

Overlap is *testable* from data (check support of $X \mid T{=}1$ vs. $X \mid T{=}0$). Weak ignorability and consistency are not testable from data alone.

5.2.6 Which Variables to Condition On: The Pre-Treatment Requirement

Before estimating anything, there is a prior graphical question: *which variables should be in X?* The answer is not "all available variables." Conditioning on the wrong variable introduces bias.

Never Condition on a Post-Treatment Variable

A variable L is *post-treatment* if $T \to L$ in the DAG. Including L in X can bias the estimated treatment effect through two mechanisms:

1. **Blocking a causal pathway.** If L is a mediator ($T \to L \to Y$), conditioning on L blocks part of the causal effect of T. The result estimates a direct effect, not the total effect.
2. **Opening a collider path.** If L is a collider ($T \to L \leftarrow U \to Y$), conditioning on L opens a previously blocked path, creating a spurious T–Y association through the unobserved U.

In both cases, including L violates Condition 1 of the back-door criterion: *X must contain no descendant of T*. The back-door criterion is not merely a sufficiency condition for identification — it is the correct filter for deciding which variables to include.

The practical rule: before running any regression, classify every candidate covariate as pre-treatment or post-treatment using the causal graph. Only pre-treatment variables satisfying the back-door criterion belong in X.

5.3 Regression Adjustment and Standardization

5.3.1 The Common Three-Step Logic

Both regression adjustment and standardization implement the same back-door formula Equation 5.5. They differ only in how they estimate $\mu(t, x) = \mathbb{E}[Y \mid T{=}t, X{=}x]$.

The Three-Step Recipe: Outcome Regression / G-Formula

1. **Estimate the outcome model.** Fit a model for $\mu(t, x) = \mathbb{E}[Y \mid T{=}t, X{=}x]$ using observed data. Any regression method may be used: OLS, logistic regression, or a flexible nonparametric estimator.
2. **Predict both potential outcomes for every unit.** For each unit i — *regardless of treatment actually received* — compute:

$$\hat{Y}_i(1) = \hat{\mu}(1, X_i), \qquad \hat{Y}_i(0) = \hat{\mu}(0, X_i).$$

   For a treated unit, $\hat{Y}_i(0)$ is the predicted outcome had that unit been assigned to control; for a control unit, $\hat{Y}_i(1)$ is the predicted outcome had that unit been treated.
3. **Average the individual treatment effect estimates.**

$$\hat{\tau}_{\text{OR}} = \frac{1}{n} \sum_{i=1}^{n} [\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)]. \tag{5.7}$$

   To estimate the ATT, average only over the n_1 treated units.

This is the *outcome regression* (OR) or *G-computation* estimator (Robins 1986). Step 3 averages out X using the empirical distribution — exactly what Equation 5.5 requires: $\tau_{\text{ATE}} = \int [\mu(1,x) - \mu(0,x)] p(x)\, dx$.

5.3.2 A Worked Example

Binary covariate X (e.g., sex), continuous outcome Y (e.g., earnings). High-X units are overrepresented in the treated arm (70% treated vs. 30% control).

	Treated ($T=1$)	Control ($T=0$)	
$X=0$	$n=30$, $\bar{Y}=42$	$n=70$, $\bar{Y}=35$	
$X=1$	$n=70$, $\bar{Y}=58$	$n=30$, $\bar{Y}=50$	
All	$n=100$, $\bar{Y}=53.2$	$n=100$, $\bar{Y}=38.5$	Unadjusted diff = 14.7

Step 1. $\hat{\mu}(1,0)=42$, $\hat{\mu}(1,1)=58$, $\hat{\mu}(0,0)=35$, $\hat{\mu}(0,1)=50$.

Step 2. Within-stratum effects: $42-35=7$ (for X=0), $58-50=8$ (for X=1).

Step 3. Averaging over the *marginal* distribution of X: 50% of the full sample has X=0, 50% has X=1:

$$\hat{\tau}_{\text{ATE}} = 7 \times 0.5 + 8 \times 0.5 = 7.5.$$

After adjusting for X, the estimated treatment effect is 7.5, not 14.7. The unadjusted comparison conflates the treatment effect with the higher baseline earnings of X=1 units who happen to be treated more often.

5.3.3 Standardization as a Special Case

When X is categorical, standardization computes $\hat{\mu}(t,x)$ as the within-cell sample mean — a fully saturated model:

$$\hat{\tau}_{\text{STD}} = \sum_{x \in \mathcal{X}} [\hat{\mu}(1,x) - \hat{\mu}(0,x)]\, \hat{p}(x). \tag{5.8}$$

This is *algebraically identical* to Equation 5.7 when X is categorical: standardization is regression adjustment with a saturated outcome model. The formula is also known as the *G-formula* (Robins 1986) and as *direct standardization* in epidemiology.

	Regression adjustment	**Standardization**
How $\hat{\mu}(t,x)$ is estimated	Parametric model: OLS, logistic, or flexible learner	Cell means: $\bar{Y}$ within $(T{=}t, X{=}x)$
Works when X is	Continuous or high-dimensional	Discrete and low-dimensional
Bias if wrong	Model misspecification	Sparse cells (some (t,x) strata empty)
Steps 2–3	Identical: predict both POs, average	Identical: predict both POs, average

5.3.4 Model Specification and What Can Go Wrong

The OR estimator is consistent if and only if $\hat{\mu}(t,x) \to \mu(t,x)$ as $n \to \infty$ — i.e., if the outcome model is correctly specified.

Misspecification bias. If the true $\mu(t,x)$ is nonlinear but a linear model is used, the estimated treatment effect absorbs the functional form error. Flexible machine learning estimators reduce this risk, at the cost of requiring cross-fitting to avoid overfitting bias (Chapter 11).

Extrapolation. Step 2 predicts $\hat{Y}_i(0)$ for treated units whose covariate values may lie outside the support of the control group. The propensity score overlap condition (Chapter 6) formalizes when extrapolation is unavoidable.

5.3.5 Regression Adjustment in Randomized Experiments: Lin (2013)

In a *randomized* experiment, outcome regression can reduce variance even though adjustment is not needed for unbiasedness. The *regression-adjusted estimator* of Lin (2013) fits the fully interacted OLS model:

$$Y_i = \alpha + \beta T_i + \gamma^\top \tilde{X}_i + \delta^\top (T_i \cdot \tilde{X}_i) + \varepsilon_i, \tag{5.9}$$

where $\tilde{X}_i = X_i - \bar{X}$ are mean-centered covariates, and takes $\hat{\beta}$ as the ATE estimate.

Why $\hat{\beta}$ equals the three-step OR estimator. The interacted regression Equation 5.9 fits arm-specific linear models with centered covariates. Under centering, the OLS intercept equals the regression-adjusted estimator of $\mathbb{E}[Y(t)]$, so $\hat{\beta} = \hat{\mu}_{1,\text{reg}} - \hat{\mu}_{0,\text{reg}} = \hat{\tau}_{\text{reg}}$.

Theorem: Design-Consistency of the Regression-Adjusted Estimator [@lin2013agnostic]

Under complete randomization, $\hat{\beta}$ from the interacted regression Equation 5.9 is consistent for τ_{ATE} regardless of whether the linear model is correctly specified.

Proof

Work in the design-based framework of **?@thm-dim**: potential outcomes fixed, only $\mathbf{T}$ random. Let $\mathbf{B}_1 = (\sum_i \mathbf{x}_i \mathbf{x}_i^\top)^{-1} \sum_i \mathbf{x}_i Y_i(1)$ be the full-population OLS coefficient (fixed under randomization), and let $e_i(1) = Y_i(1) - \mathbf{x}_i^\top \mathbf{B}_1$. By normal equations, $\sum_i e_i(1) = 0$ and $\frac{1}{N}\sum_i \mathbf{x}_i^\top \mathbf{B}_1 = \bar{Y}(1)$.
Linearization. A short algebraic manipulation using normal equations gives:

$$\hat{\mu}_{1,\text{reg}} = \bar{Y}(1) + \frac{1}{N_1} \sum_{i=1}^{N} T_i e_i(1) + O_p(N^{-1}). \tag{5.10}$$

The remainder term R_N is $O_p(N^{-1})$ because the covariate-balance gap is $O_p(N^{-1/2})$ by design, and $\hat{\beta}_1 - \mathbf{B}_1 = O_p(N^{-1/2})$ by standard within-arm OLS theory.
Consistency. Under CRE, each unit has $\Pr(T_i = 1 \mid \mathcal{F}_N) = N_1/N$, so $\mathbb{E}[N_1^{-1}\sum_i T_i e_i(1) \mid \mathcal{F}_N] = 0$ by $\sum_i e_i(1) = 0$. The conditional variance is $O(N^{-1})$ by finite-population sampling theory; by Chebyshev, $N_1^{-1}\sum_i T_i e_i(1) \to_p 0$. By Equation 5.10, $\hat{\mu}_{1,\text{reg}} - \bar{Y}(1) \to_p 0$. Symmetrically, $\hat{\mu}_{0,\text{reg}} - \bar{Y}(0) \to_p 0$. Therefore $\hat{\beta} \to_p \bar{Y}(1) - \bar{Y}(0) = \tau_{\text{ATE}}$.
Crucially, the probability limit of $\hat{\beta}_1$ never enters: consistency comes from the randomization distribution, not from correctness of the linear model. □

Byproduct: variance reduction. Equation Equation 5.10 shows that the regression estimator is approximately $\bar{\tau} + N_1^{-1}\sum_i T_i e_i(1) - N_0^{-1}\sum_i (1-T_i) e_i(0)$. The variance reduction relative to DIM comes from replacing raw outcomes by residuals: when the linear model explains a meaningful share of $Y_i(t)$'s variation, $S^2_{e(t)} < S^2_{Y(t)}$ and the design variance shrinks.

Remark: The Role of the Regression Model in a Randomized Experiment

In a randomized experiment, the regression model serves *one purpose only*: variance reduction. It does not contribute to unbiasedness. DIM is already design-consistent; adding covariates reduces $\text{Var}(\hat{\tau})$ by absorbing residual variation in Y unrelated to T. If the model fits well, the residual variance shrinks and the estimator becomes more precise. If misspecified or covariates are weakly predictive, the variance reduction is small or zero, but *no bias is introduced*. This stands in sharp contrast to the observational setting, where the regression model carries a double burden: it must both adjust for confounding (unbiasedness) and fit the outcome surface (efficiency). Misspecification in an observational study threatens *both* properties simultaneously.

Remark: Precedence in Survey Sampling

The design-consistency result in **?@thm-lin** — model-agnostic consistency under the randomization distribution — was established in the survey sampling literature decades before Lin (2013). Isaki and Fuller (1982) proved this for the generalized regression (GREG) estimator under general probability sampling designs. A CRE is structurally equivalent to simple random sampling from the finite population of potential outcomes, so their result directly implies **?@thm-lin**. Deville and Särndal (1992) extended this to calibration estimators. Lin (2013)'s contribution was restating and proving this within the Neyman–Rubin framework for the experimental causal inference audience.

5.4 Stratification

Stratification (subclassification) is a non-parametric implementation of the back-door formula. Rather than modeling $\mathbb{E}[Y \mid T, X]$, the analyst divides the sample into *strata* — subgroups with similar values of X — and estimates the treatment effect within each stratum.

Within stratum $\mathcal{S}_k$, the treated and control units have similar covariate distributions. If the stratum is narrow enough, ignorability holds approximately, and the within-stratum DIM:

$$\hat{\tau}_k = \bar{Y}_{1,k} - \bar{Y}_{0,k}$$

is approximately unbiased for the stratum-specific ATE. The overall ATE is estimated as:

$$\hat{\tau}_{\text{STRAT}} = \sum_{k=1}^{K} \hat{\tau}_k \cdot \frac{n_k}{n}. \tag{5.11}$$

Cochran (1968) showed that with $K = 5$ equal-size strata, roughly 90% of the bias from a single continuous confounder is removed; with $K = 10$, over 95%.

Remark: Stratification, Standardization, and Survey Sampling Terminology

When X is categorical, the stratified estimator Equation 5.11 and the standardization estimator Equation 5.8 are algebraically identical: both compute within-cell treatment effect estimates and aggregate with marginal weights n_k/n. The names reflect disciplinary tradition.
When X is continuous, stratification estimates $\hat{\mu}(t, x)$ by a piecewise-constant step function; regression fits a smooth model. In the survey sampling literature, *stratification* (design-stage: divide before sampling) and *post-stratification* (analysis-stage: reweight after sampling) are distinct. Standardization in causal inference corresponds to post-stratification.

Limitations. Stratification faces the *curse of dimensionality*: with p binary covariates, there are 2^p cells, many empty in practice. Two solutions: regression adjustment, which imposes parametric structure on $\hat{\mu}(t, x)$; and propensity score stratification (Chapter 6), which collapses multivariate X to a scalar $\pi(X) = P(T{=}1 \mid X)$.

5.5 Simpson's Paradox

Simpson's paradox — the reversal of an association when conditioning on a third variable — was introduced in Chapter 1. This chapter's development gives it a second layer of interpretation. The back-door formula used in Chapter 1 is precisely the standardization estimator Equation 5.8: within-stratum conditional means averaged over the *marginal* distribution of the confounder rather than its treatment-conditional distribution. The pooled association does not do this — it weights strata by $P(X \mid T)$ instead of $P(X)$.

5.5.1 When Should You *Not* Condition on X?

Simpson's paradox has a mirror image: conditioning can *create* a spurious association or make a true causal effect disappear. This occurs when X is a mediator or a collider. The back-door criterion gives the correct prescription: include X only if it blocks back-door paths *without* opening collider paths.

Conditioning on a Mediator or Collider

Adjusting for a post-treatment variable X that lies on a causal pathway $T \to X \to Y$ blocks part of the causal effect, producing a downward-biased estimate of the total effect. Adjusting for a collider X with $T \to X \leftarrow Y$ opens a spurious association that does not exist in the population. Neither mistake is detectable from the data alone; the DAG is essential.

5.6 Lab: Simulation Study of the Outcome Regression Estimator

This simulation compares the linear and local-linear OR estimators across two designs and illustrates the bias-variance tradeoff.

Data-generating process. $X \sim \text{Uniform}(0,1)$, $\varepsilon \sim \mathcal{N}(0,1)$ independent of X. Potential outcomes:

$$Y(t) = t + 3(1+t)X^2 + \varepsilon, \qquad t \in \{0,1\}. \tag{5.12}$$

The conditional ATE is $\tau(x) = 1 + 3x^2$, so $\tau_{\text{ATE}} = 1 + 3\mathbb{E}[X^2] = 1 + 1 = 2$.

Two assignment mechanisms:

- **CRD**: $T \sim \text{Bern}(0.5)$, independent of X.
- **Observational**: $T \mid X \sim \text{Bern}(\text{expit}(-2 + 5X))$, giving $\pi(0) \approx 0.12$, $\pi(0.5) \approx 0.62$, $\pi(1) \approx 0.95$. Strong confounding: high-X units are nearly always treated.

Strong ignorability holds in both: under CRD by design; under the observational mechanism because T depends only on X through a known stochastic function, leaving no unmeasured variable affecting both T and Y.

Estimators. *Linear OR*: within each arm, fit OLS of Y on X (misspecified: true conditional mean is quadratic with a TX^2 interaction). *Local-linear OR*: within each arm, fit local linear regression with Gaussian kernel, bandwidth $h = 0.5 \cdot n^{-1/5}$.

Results ($n = 1000$, $B = 2000$ replications, seed 2024):

Design	Estimator	Mean	Bias	SD	RMSE
CRD	Linear OR	2.0013	+0.0013	0.0679	0.0678
CRD	Local-linear OR	2.0333	+0.0333	0.0659	0.0738
Observational	Linear OR	1.9602	−0.0398	0.0851	0.0939
Observational	Local-linear OR	2.0240	+0.0240	0.0923	0.0953

Lesson 1: Under CRD, the linear OR is unbiased regardless of model specification. This confirms **?@thm-lin**: under complete randomization, the linear OR converges to τ_{ATE} even with a wrong outcome model. The randomization distribution, not the model, does the identification work.

Lesson 2: Under an observational design, the misspecified linear OR is biased. The bias grows from +0.0013 under CRD to −0.0398 under the observational design. The omitted X^2 and TX^2 terms cause $\hat{\mu}(t,x)$ to misrepresent the outcome surface; because high-X units are predominantly treated, the misfit is systematically amplified in the direction of confounding. Under CRD, the balanced assignment averages out the same misfit.

Lesson 3: The local-linear OR nearly eliminates bias at the cost of higher variance. Under the observational design, local-linear OR reduces bias from −0.0398 to +0.0240 but its SD is 0.0923 vs. 0.0851 for the linear model. The RMSE comparison (0.0939 vs. 0.0953) favors the linear estimator in MSE, even though it is biased — the classic bias-variance tradeoff.

Model Misspecification and Confounding Interact

The linear OR bias is +0.0013 under CRD and −0.0398 under the observational design. The misspecification is identical in both cases — the difference comes entirely from the assignment mechanism. Under CRD, $T \perp\!\!\!\perp X$, so X within each arm is Uniform$(0,1)$. The population

OLS pointwise error $b(x) = -\frac{1}{2} + 3x - 3x^2$ integrates to exactly zero over $U(0,1)$: $\mathbb{E}[b(X)] = \int_0^1 (-\frac{1}{2} + 3x - 3x^2)\,dx = 0$. This cancellation is structural: the population OLS line is the linear projection of $\tau(X)$ onto $\{1, X\}$ under $U(0,1)$, and a linear projection always preserves the mean. Under the observational design, the within-arm distributions of X are distorted by the propensity score, so the same cancellation fails.

5.7 Chapter Summary

Estimand	Identification	Key assumption	Method
ATE under CRD	$\mathbb{E}[Y(t)] = \mathbb{E}[Y \mid T{=}t]$	Randomization protocol	DIM, regression-adjusted DIM
ATE under observability	Equation 5.5	Ignorability + overlap	OR, standardization, stratification
ATT	Equation 5.6	Same, one-sided overlap for $t = 0$	Same methods, weighted over treated
Propensity score (Ch. 6)	Reduces X to scalar $\pi(X)$	Correctly specified PS model	IPW, matching, PS stratification

1. **Randomization as graph surgery.** A randomized experiment removes all arrows into T, so $f(y \mid \text{do}(T{=}t)) = f(y \mid T{=}t)$ and DIM is unbiased for the ATE without covariate adjustment.
2. **Ignorability is the key assumption.** Under unconfoundedness $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$, the back-door adjustment formula Equation 5.5 identifies the ATE. This assumption is substantive, design-determined in RCTs, and untestable from observational data.
3. **Three estimation strategies.** Regression adjustment (G-formula), standardization, and stratification all implement Equation 5.5 via different modeling choices. Under a randomized experiment, regression adjustment is design-consistent for the ATE even when the outcome model is misspecified (**?@thm-lin**); in observational studies, it is consistent only when the outcome model is.
4. **Simpson's paradox.** Pooled associations can reverse within subgroups when a confounder determines treatment selection. The correct causal estimate requires standardization over the *marginal* covariate distribution, guided by the back-door criterion.
5. **The propensity score dimension reduction.** When X is high-dimensional, direct stratification or cell-by-cell standardization breaks down. The propensity score $\pi(X)$ provides a scalar sufficient statistic for adjustment. Chapter 6 develops the theory and estimation methods.

5.8 Problems

1. Randomization and the do-calculus. Let the observational DAG be $\{U \to T, U \to Y, X \to T, T \to Y\}$ with U unobserved.

(a) Identify all back-door paths from T to Y.
(b) Does X satisfy the back-door criterion? Does U? Explain.
(c) Now suppose treatment is randomized. Draw the modified DAG and identify all back-door paths. Show that $f(y \mid \text{do}(T{=}t)) = f(y \mid T{=}t)$ holds using Rule 2 of the do-calculus. (*Hint*: identify the appropriate (X, Z, W) instantiation, construct $\mathcal{G}_{\underline{T}}$, and verify $(Y \perp\!\!\!\perp T)_{\mathcal{G}_{\underline{T}}}$.)
(d) Under randomization, is covariate adjustment on X necessary for unbiasedness? Is it ever beneficial? Explain.

2. Ignorability and the back-door criterion. Consider the DAG $\{X \to T, X \to Y, T \to Y, U \to Y\}$ with U unobserved.

(a) Does $\{X\}$ satisfy the back-door criterion? Write out the identifying formula for τ_{ATE}.
(b) Now add the edge $U \to T$ to the DAG. Does $\{X\}$ still satisfy the back-door criterion? What does the criterion require when both X and U affect T?
(c) Explain in words what "unconfoundedness given X" means about the role of U in the data-generating process.

3. Standardization. A study of job training (T) and earnings (Y, in thousands) produces the following cell means, with $P(X{=}0) = 0.4$ and $P(X{=}1) = 0.6$:

X	$\hat{\mu}(1, x)$	$\hat{\mu}(0, x)$	n_x/n
0	28	22	0.4
1	35	31	0.6

(a) Compute $\hat{\tau}_{\text{ATE}}$ using the standardization formula Equation 5.8.
(b) Compute $\hat{\tau}_{\text{ATT}}$, given that all treated units come from $X{=}1$ — i.e., $P(X{=}1 \mid T{=}1) = 1$.
(c) The unadjusted difference in means is $33 - 25 = 8$. Compare to your answers in (a) and (b) and explain the discrepancy.

4. Stratification and Cochran's rule. You have a binary confounder $X \in \{0, 1\}$ and form two strata. Within stratum $X{=}0$: $n_0 = 600$, $\bar{Y}_1 = 10$, $\bar{Y}_0 = 8$. Within stratum $X{=}1$: $n_1 = 400$, $\bar{Y}_1 = 15$, $\bar{Y}_0 = 12$.

(a) Compute the stratified ATE estimator $\hat{\tau}_{\text{STRAT}}$ via Equation 5.11.
(b) Suppose an unadjusted DIM gives $\hat{\tau}_{\text{DIM}} = 5.5$. Explain why the two estimates differ and which is the appropriate causal estimate.
(c) Describe one limitation that would arise if X were a continuous variable with 15 dimensions.

5. Simpson's paradox. A hospital reports that ICU patients ($T{=}1$) have higher mortality than others ($T{=}0$): $P(Y{=}1 \mid T{=}1) = 0.30$ vs. $P(Y{=}1 \mid T{=}0) = 0.10$.

(a) Construct a numerical example (a $2 \times 2 \times 2$ table with disease severity X as the confounder) consistent with the pooled numbers yet showing ICU admission *reduces* mortality within both severity strata.
(b) Compute the standardized ATE using the marginal distribution of X.
(c) Draw the DAG. Identify the back-door path that the pooled comparison fails to block.
(d) A colleague argues the pooled statistic is the right answer since the hospital treats patients with both mild and severe illness. Explain, using potential outcomes notation, why this argument is incorrect.

Chapter 6

Propensity Score Methods

Learning Objectives

By the end of this chapter, students should be able to:

1. Define the propensity score $\pi(X) = P(T{=}1 \mid X)$ and explain why it is a balancing score.
2. State and prove the Balancing Property ($T \perp\!\!\!\perp X \mid \pi(X)$), the Coarsest Balancing Score Lemma, and the Propensity Score Theorem.
3. Describe standard methods for estimating the propensity score and the practical pitfalls of each.
4. Derive the IPW identification formula for the ATE and explain the role of the Horvitz–Thompson representation.
5. Distinguish the IPW estimator (targets ATE) from the nearest-neighbor matching estimator (targets ATT).
6. Distinguish positivity (weak overlap) from strong overlap, describe the practical consequences of near-violations, and apply trimming strategies.
7. Articulate the fundamental limitation of propensity score methods — the unconfoundedness assumption — and explain why this motivates instrumental variables (Chapter 7).

6.1 Motivation: The Curse of Dimensionality

Chapter 5 established that under strong ignorability, the ATE is identified by the back-door adjustment formula, and developed three estimation strategies: regression adjustment, standardization, and stratification. All three require conditioning on a covariate vector X. When X is low-dimensional, this is feasible. When X has many components, it is not.

The problem. Stratification requires cells $\{X = x\}$ to contain both treated and control units. With p binary covariates, there are 2^p cells. With $p = 10$, that is 1024 cells — far more than most datasets can support. Regression adjustment avoids the cell-count problem but requires a correctly specified model for $\mathbb{E}[Y \mid T, X]$, which becomes harder to specify reliably as p grows.

The solution. Rosenbaum and Rubin (1983) showed that it is sufficient to condition on a single scalar function of X: the *propensity score* $\pi(X) = P(T{=}1 \mid X)$. The propensity score reduces the adjustment dimension without changing the identified estimand, provided strong ignorability and positivity hold.

6.2 The Propensity Score and Its Balancing Properties

6.2.1 Balancing Scores and the Propensity Score

Definition: Balancing Score [@rosenbaum1983central]

A function $b(X)$ is called a **balancing score** if $T \perp\!\!\!\perp X \mid b(X)$.

The full covariate vector X is trivially a balancing score. The propensity score is the *coarsest* balancing score, providing the greatest dimension reduction without sacrificing identification.

Definition: Propensity Score [@rosenbaum1983central]

For a binary treatment $T \in \{0, 1\}$ and observed covariates X, the **propensity score** is $\pi(X) = P(T{=}1 \mid X)$.

In a randomized experiment, $\pi(X) = 0.5$ for all units. In an observational study, $\pi(X)$ varies with X and must be estimated from data.

Connection to strong ignorability. Under strong ignorability $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$, the potential outcomes carry no further information about treatment assignment once X is given: $P(T{=}1 \mid X, Y(0), Y(1)) = P(T{=}1 \mid X) = \pi(X)$. This equality is a *consequence* of the ignorability assumption, not part of the definition of $\pi(X)$.

Theorem: The Propensity Score Is a Balancing Score [@rosenbaum1983central]

$T \perp\!\!\!\perp X \mid \pi(X)$.

Proof

It suffices to show $P(T{=}1 \mid X, \pi(X)) = P(T{=}1 \mid \pi(X))$. Since $\pi(X)$ is a deterministic function of X, $P(T{=}1 \mid X, \pi(X)) = P(T{=}1 \mid X) = \pi(X)$. For the right-hand side, the law of iterated expectations gives:

$$P(T{=}1 \mid \pi(X)) = \mathbb{E}[P(T{=}1 \mid X) \mid \pi(X)] = \mathbb{E}[\pi(X) \mid \pi(X)] = \pi(X). \quad \square$$

Remark: Covariate Balance and Design-Stage Analysis

The balancing property implies $f(X \mid T{=}1, \pi(X)) = f(X \mid T{=}0, \pi(X))$ a.s. Imbens and Rubin (2015) call this *design before analysis*: covariate balance can be assessed by stratifying on a discretized $\hat{\pi}(X)$ and comparing covariate distributions within strata — without consulting the outcome Y, making the diagnostic immune to outcome-fishing.

Remark: Observed Covariates Only

The Balancing Property is a statement about the *observed* covariate distribution. It does *not* say anything about unobserved confounders U: if U affects both T and Y and is not captured by X, the balancing property fails to close the back-door path through U.

6.2.2 The Propensity Score Theorem

Lemma: The Propensity Score Is the Coarsest Balancing Score [@rosenbaum1983central]

If $b(X)$ is any balancing score — that is, $T \perp\!\!\!\perp X \mid b(X)$ — then $\pi(X)$ is a function of $b(X)$: $\pi(X) = g(b(X))$ for some measurable g.

Proof

Since $b(X)$ is a balancing score, $P(T{=}1 \mid X, b(X)) = P(T{=}1 \mid b(X))$. Since $b(X)$ is a function of X: $\pi(X) = P(T{=}1 \mid X) = P(T{=}1 \mid X, b(X)) = P(T{=}1 \mid b(X))$. Hence $\pi(X)$ is a measurable function of $b(X)$ alone. □

The lemma states that $\pi(X)$ is the smallest sufficient statistic for T among functions of X: any other balancing score retains at least as much information about X.

Theorem: Propensity Score Theorem [@rosenbaum1983central]

Suppose strong ignorability holds: $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$ and $0 < \pi(X) < 1$ a.s. Then for any balancing score $b(X)$:

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid b(X).$$

In particular, $(Y(0), Y(1)) \perp\!\!\!\perp T \mid \pi(X)$.

Proof

We prove the result for $b(X) = \pi(X)$. For any $t \in \{0, 1\}$:

$$P(T{=}1 \mid Y(t), \pi(X)) = \mathbb{E}[P(T{=}1 \mid Y(t), X) \mid Y(t), \pi(X)] = \mathbb{E}[P(T{=}1 \mid X) \mid Y(t), \pi(X)] = \mathbb{E}[\pi(X) \mid Y(t), \pi(X)] = \pi(X)$$

where the second equality uses ignorability ($T \perp\!\!\!\perp Y(t) \mid X$). Hence $T \perp\!\!\!\perp Y(t) \mid \pi(X)$. By the Coarsest Balancing Score Lemma, the result extends to any $b(X)$. □

Remark: Identification via the Propensity Score

?@thm-ps gives an alternative identification formula: $\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y \mid \pi(X), T{=}t]]$. A common misconception is that $\mathbb{E}[Y \mid \pi(X), T{=}t]$ and $\mathbb{E}[Y \mid X, T{=}t]$ are equal pointwise — they are not. The theorem guarantees only that their marginal expectations agree: $\mathbb{E}[\mathbb{E}[Y \mid \pi(X), T{=}t]] = \mathbb{E}[\mathbb{E}[Y \mid X, T{=}t]] = \mathbb{E}[Y(t)]$.

The Dimension Reduction

?@thm-ps reduces the curse of dimensionality from a $\dim(X)$-dimensional problem to a one-dimensional one:

$$\tau_{\text{ATE}} = \mathbb{E}[\mathbb{E}[Y \mid T{=}1, \pi(X)] - \mathbb{E}[Y \mid T{=}0, \pi(X)]].$$

6.3 Estimation of the Propensity Score

Logistic regression. The classical approach models $\text{logit}(\pi(X)) = X^\top \beta$ and estimates β by maximum likelihood. The fitted values $\hat{\pi}(X_i) = \sigma(X_i^\top \hat{\beta})$ are used in place of the true propensity score.

Machine learning estimators. When X is high-dimensional or the true propensity score is nonlinear, machine learning methods — gradient boosted trees, random forests, regularized logistic regression — can estimate $\pi(X)$ more flexibly. A critical issue is *overfitting*: an estimator that perfectly separates treated and control units in-sample will produce estimated propensity scores near 0 or 1 everywhere, destroying overlap. Cross-fitting (Chapter 11) addresses this by estimating nuisance functions on held-out folds.

Propensity Score Estimation Is a Nuisance, Not the Goal

The propensity score is estimated to construct a good estimator of τ_{ATE}. Balancing tests diagnose whether the estimated score has achieved covariate balance, but they do *not* test whether unobserved confounders are balanced.

6.4 Matching on the Propensity Score

The propensity score motivates *matching*: for each treated unit i, find a control unit j with $\hat{\pi}(X_j) \approx \hat{\pi}(X_i)$, and use Y_j as a local estimate of $\mathbb{E}[Y(0) \mid \pi(X) = \pi(X_i)]$ (Abadie and Imbens 2006, 2016). The matched control outcome is not a substitute for the individual counterfactual $Y_i(0)$; it approximates the conditional mean:

$$\mathbb{E}[Y(0) \mid \pi(X) = \pi(X_i)] = \mathbb{E}[Y(0) \mid \pi(X) = \pi(X_i), T{=}0] = \mathbb{E}[Y \mid \pi(X) = \pi(X_i), T{=}0],$$

where the first equality uses **?@thm-ps** and the second uses consistency. Averaging across treated units recovers the ATT:

$$\hat{\tau}_{\text{ATT}} = \frac{1}{n_1} \sum_{i:\, T_i=1} [Y_i - Y_{\hat{\jmath}(i)}], \tag{6.1}$$

where $\hat{\jmath}(i)$ is the matched control index.

One-to-one nearest-neighbor matching. For each treated unit i:

$$\hat{\jmath}(i) = \arg\min_{j:\, T_j=0} |\hat{\pi}(X_i) - \hat{\pi}(X_j)|.$$

Matching *with replacement* reduces bias but increases variance.

Caliper matching. Restricts matches to pairs within a maximum distance δ, discarding treated units with no close match. This restricts the estimand to a subpopulation with good overlap.

> Remark: Matching as Nonparametric Regression on $\pi(X)$
>
> Propensity score matching implicitly fits a nonparametric regression of Y on $\pi(X)$ separately within each treatment arm. By **?@thm-ps**, $\mathbb{E}[Y(t) \mid \pi(X) = p]$ is identified from observed data for each t. Nearest-neighbor matching approximates this by averaging outcomes of close matches — local constant regression on the scalar propensity score. The dimension reduction from **?@thm-ps** is what makes nonparametric estimation feasible.

6.5 Inverse Probability Weighting

6.5.1 The IPW Identification Formula

Under strong ignorability, the ATE has an *inverse probability weighting* (IPW) representation:

$$\tau_{\text{ATE}} = \mathbb{E}\left[\frac{T \cdot Y}{\pi(X)}\right] - \mathbb{E}\left[\frac{(1-T) \cdot Y}{1-\pi(X)}\right].$$

> Theorem: IPW Identification
>
> Under strong ignorability ($(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$ and $0 < \pi(X) < 1$): $\mathbb{E}[TY/\pi(X)] = \mathbb{E}[Y(1)]$.

> Proof
>
> $$\mathbb{E}\left[\frac{TY}{\pi(X)}\right] = \mathbb{E}\left[\frac{TY(1)}{\pi(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{TY(1)}{\pi(X)} \,\middle|\, X\right]\right] = \mathbb{E}\left[\frac{Y(1)}{\pi(X)}\,\mathbb{E}[T \mid X]\right] = \mathbb{E}\left[\frac{Y(1)}{\pi(X)} \cdot \pi(X)\right] = \mathbb{E}[Y(1)].$$
>
> The first equality uses consistency ($Y = Y(1)$ when $T{=}1$); the third uses ignorability ($Y(1) \perp\!\!\!\perp T \mid X$). □

Remark: Scope of the Ignorability Assumption

The proof invokes only the marginal independence $Y(1) \perp\!\!\!\perp T \mid X$, not the full joint $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$. An analogous proof for the control term uses only $Y(0) \perp\!\!\!\perp T \mid X$. The joint independence is used in this chapter for uniformity with Chapter 5 and because it is needed for estimands involving the joint distribution of potential outcomes.

6.5.2 Horvitz–Thompson and Hájek Estimators

The **Horvitz–Thompson (HT)** IPW estimator:

$$\hat{\tau}_{\text{HT}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1 - T_i) Y_i}{1 - \hat{\pi}(X_i)} \right]. \tag{6.2}$$

The **Hájek (self-normalized)** estimator:

$$\hat{\tau}_{\text{HJ}} = \frac{\sum_i T_i Y_i / \hat{\pi}(X_i)}{\sum_i T_i / \hat{\pi}(X_i)} - \frac{\sum_i (1 - T_i) Y_i / (1 - \hat{\pi}(X_i))}{\sum_i (1 - T_i) / (1 - \hat{\pi}(X_i))}. \tag{6.3}$$

The Hájek estimator caps each unit's effective weight at its share of the total within its arm, suppressing the influence of extreme weights. The efficient doubly robust AIPW estimator of Chapter 10 combines the outcome model and the propensity score; under correct nuisance specification it achieves higher efficiency than either pure IPW or pure outcome regression.

ATT estimation. The ATT re-weights the *control arm* to look like the treated population:

$$\hat{\tau}_{\text{HT,ATT}} = \frac{1}{n_1} \sum_{i:\, T_i = 1} Y_i - \frac{1}{n_1} \sum_{i:\, T_i = 0} \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} Y_i, \tag{6.4}$$

$$\hat{\tau}_{\text{HJ,ATT}} = \frac{1}{n_1} \sum_{i:\, T_i = 1} Y_i - \frac{\sum_{i:\, T_i = 0} \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} Y_i}{\sum_{i:\, T_i = 0} \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)}}. \tag{6.5}$$

Each control unit receives weight proportional to its *odds of treatment* $\hat{\pi}/(1 - \hat{\pi})$. For a control unit with $\hat{\pi}(X_i) = 0.95$ the weight is 19; combined with a large baseline outcome this produces extreme variance.

6.6 Overlap and Positivity

6.6.1 Positivity and Strong Overlap

Definition: Positivity (Weak Overlap)

The **positivity** condition requires $0 < \pi(X) < 1$ almost surely. Every unit has a positive probability of receiving either treatment or control, regardless of its covariate values.

Together with unconfoundedness, positivity constitutes the strong ignorability of Rosenbaum and Rubin (1983). Positivity is the condition needed for *identification*.

Definition: Strong Overlap

The **strong overlap** condition requires $c \leq \pi(X) \leq 1 - c$ a.s. for some $c \in (0, 1/2)$.

Strong overlap bounds the inverse weights uniformly away from infinity. It is the condition under which $\sqrt{n}$-consistent, asymptotically normal inference for the ATE goes through (Chapter 11). Positivity alone permits identification but does not guarantee stable inference: when $\pi(X)$ approaches 0 or 1, IPW weights blow up even though the parameter is technically identified (Khan and Tamer 2010).

6.6.2 Practical Consequences of Near-Violations

Near-Overlap Problems

When $\hat{\pi}(X_i)$ is close to 0 or 1, the IPW weight is very large. A small number of units with extreme weights can dominate the estimator, increasing variance dramatically.

6.6.3 Trimming Strategies

Trimming by propensity score. Restrict to units with $\eta \leq \pi(X) \leq 1-\eta$ for small $\eta > 0$. The estimand changes to:

$$\tau_{\text{trim}} = \mathbb{E}[Y(1) - Y(0) \mid \eta \leq \pi(X) \leq 1 - \eta].$$

Crump et al. (2009) rule. Crump et al. (2009) derive the optimal trimming threshold minimizing the asymptotic variance of the trimmed ATE estimator. The rule $\eta = 0.1$ is a common default.

6.6.4 Re-targeting the Estimand

When overlap fails, an alternative to trimming is to change the *estimand*. The ATT requires only $P(T{=}0 \mid X{=}x) > 0$ for x in the support of $X \mid T{=}1$, i.e., $\pi(X) < 1$ a.s. on the treated support. The ATT tolerates regions with $\pi(x) = 0$ (no treated units there) but not regions inside the treated support where $\pi(x) = 1$. The ATC is symmetric. When overlap fails where $\pi(X) \approx 0$, re-target to the ATT; when overlap fails where $\pi(X) \approx 1$, re-target to the ATC.

6.7 Lab: Simulation Study of IPW and Matching Estimators

This lab compares five estimators on a five-covariate DGP with treatment-effect heterogeneity. With five continuous confounders, direct stratification is infeasible, making propensity-score methods the natural choice.

6.7.1 Part 1: Correctly Specified Propensity Score

DGP for Lab 6

$X_j \overset{\text{i.i.d.}}{\sim} N(0,1)$ for $j = 1, \ldots, 5$. All five covariates are genuine confounders. True propensity score:

$$\pi(X) = \text{expit}(-0.5 + 0.8X_1 + 0.5X_2 - 0.3X_3 + 0.2X_4 + 0.1X_5), \tag{6.6}$$

giving $P(T{=}1) \approx 0.40$. Potential outcomes:

$$Y(t) = (1 + 0.5X_1) \cdot t + 8 + 0.8X_1 + 0.5X_2 + 0.4X_3 + 0.3X_4 + 0.2X_5 + \varepsilon, \quad \varepsilon \sim N(0,1). \tag{6.7}$$

The CATE is $\tau(X) = 1 + 0.5X_1$, so $\tau_{\text{ATE}} = 1.000$ (exact). Because $\pi(X)$ is increasing in X_1, treated units have $\mathbb{E}[X_1 \mid T{=}1] > 0$, giving $\tau_{\text{ATT}} \approx 1.199$ (oracle, $n = 10^7$). The gap ≈ 0.199 reflects selection: units most likely to be treated also tend to benefit more.

Estimators. Five estimators are compared under both the true and estimated propensity score. The estimated score uses logistic regression of T on $(X_1, \ldots, X_5)$ — correctly specified since the true logit is linear. The five estimators are: HT-ATE Equation 6.2, HJ-ATE Equation 6.3 (both targeting ATE); HT-ATT Equation 6.4, HJ-ATT Equation 6.5, and NNM Equation 6.1 (all targeting ATT). The large baseline mean $\mathbb{E}[Y(0)] = 8$ amplifies the consequences of extreme IPW weights, making Hájek normalization strongly recommended.

Results ($n = 1{,}000$, $B = 2{,}000$ replications, seed 42). Bias relative to each estimator's own target.

PS	Estimator	Estimand	Mean	Bias	SD	RMSE
Known	HT-ATE	ATE	1.008	+0.008	0.626	0.626

PS	Estimator	Estimand	Mean	Bias	SD	RMSE
Known	HJ-ATE	ATE	1.003	+0.003	0.145	0.145
Known	HT-ATT	ATT	1.184	−0.016	0.725	0.725
Known	HJ-ATT	ATT	1.203	+0.003	0.131	0.131
Known	NNM	ATT	1.207	+0.008	0.130	0.130
Estimated	HT-ATE	ATE	0.998	−0.002	0.194	0.194
Estimated	HJ-ATE	ATE	1.002	+0.002	0.102	0.103
Estimated	HT-ATT	ATT	1.202	+0.002	0.251	0.251
Estimated	HJ-ATT	ATT	1.202	+0.002	0.101	0.101
Estimated	NNM	ATT	1.205	+0.006	0.121	0.121

Lesson 1: Hájek normalization is strongly recommended when outcomes are non-centered. With the known PS, HT-ATE achieves SD = 0.626, while HJ-ATE achieves SD = 0.145 — a **4.3-fold reduction**. A unit with $\pi(X_i) = 0.05$ receives raw weight 20; multiplied by an outcome near $\mathbb{E}[Y(0)] = 8$, its contribution is of order 160.

Lesson 2: HT-ATT is even more unstable; Hájek normalization is equally beneficial. HT-ATT SD = 0.725. A control unit with $\hat{\pi}(X_i) = 0.95$ receives odds-ratio weight 19. HJ-ATT reduces SD to 0.131 — a **5.5-fold reduction**.

Lesson 3: HJ-ATT and NNM converge to the same target with nearly identical efficiency. Under the known PS, HJ-ATT (SD = 0.131) and NNM (SD = 0.130) are virtually indistinguishable — a clear "two routes, one estimand" demonstration.

Lesson 4: The estimand gap between ATE and ATT is large and clearly revealed. The ATE estimators converge to 1.000; the ATT estimators converge to ≈ 1.199. The 0.199 gap is not bias. A researcher who applies NNM and reports against the ATE benchmark would conclude the estimator has 20% bias; it is actually unbiased for the correct target.

Lesson 5: Estimated PS collapses HT variance. HT-ATE SD falls from 0.626 to 0.194 (69% reduction); shrinkage of fitted logistic probabilities toward the sample mean trims extreme weights automatically. HJ-ATT with estimated PS achieves SD = 0.101 — beating NNM (SD = 0.121).

> The Estimand Must Be Chosen Before the Estimator
>
> In this DGP, $\tau_{\text{ATE}} = 1.000$ and $\tau_{\text{ATT}} \approx 1.199$. The gap arises because treatment selection is correlated with the individual treatment effect: high-X_1 units are simultaneously more likely to be treated and more likely to benefit (Imbens and Rubin 2015). Choosing among estimators should be driven by the policy question (ATE or ATT?), not by which produces the preferred point estimate.

6.7.2 Part 2: Robustness to PS Model Misspecification

Modified DGP. Outcome model unchanged. True PS now contains a quadratic term:

$$\pi^*(X) = \text{expit}(-0.5 + 0.8X_1 + 0.5X_1^2 + 0.2X_4 + 0.1X_5), \tag{6.8}$$

giving a U-shaped propensity surface. $\tau_{\text{ATE}} = 1.000$ (unchanged); $\tau^*_{\text{ATT}} \approx 1.152$ (oracle). The *estimated* PS is still a linear logistic regression on $(X_1, \ldots, X_5)$ — misspecified by omitting X_1^2.

Results ($n = 1{,}000$, $B = 2{,}000$, seed 42):

PS	Estimator	Estimand	Mean	Bias	SD	RMSE
True	HT-ATE	ATE	1.028	+0.028	0.824	0.825
True	HJ-ATE	ATE	1.013	+0.013	0.152	0.152
True	HT-ATT	ATT	1.186	+0.034	1.333	1.333
True	HJ-ATT	ATT	1.183	+0.031	0.212	0.215
True	NNM	ATT	1.178	+0.026	0.175	0.177
Misspecified	HT-ATE	ATE	1.478	+0.478	0.147	0.501
Misspecified	HJ-ATE	ATE	0.861	−0.139	0.087	0.164

PS	Estimator	Estimand	Mean	Bias	SD	RMSE
Misspecified	HT-ATT	ATT	1.780	+0.628	0.162	0.649
Misspecified	HJ-ATT	ATT	1.306	+0.154	0.081	0.174
Misspecified	NNM	ATT	1.172	+0.020	0.138	0.139

Lesson 6: PS misspecification makes all IPW estimators inconsistent; Hájek cannot fix this. HT-ATE bias is +0.478 and HJ-ATE bias is −0.139 — *opposite signs*. Hájek removes the *level error* (weight sums drifting from 1) but not the *shape error* (misweighting of the population). With Monte Carlo means $w_T \approx 1.045$ and $w_C \approx 0.972$, the extra HT bias is approximately $(w_T - 1)\mathbb{E}[Y(1)] - (w_C - 1)\mathbb{E}[Y(0)] \approx 0.045 \times 9 + 0.028 \times 8 \approx +0.62$, accounting for the entire gap $+0.478 - (-0.139) = +0.617$.

Lesson 7: In this simulation, matching is less sensitive to PS misspecification. NNM bias barely changes: +0.026 under the true PS versus +0.020 under misspecification. NNM does not require the PS model to be correct — it only requires that the estimated score roughly *orders* units so matched pairs are approximately balanced on true confounders. The linear logistic model, though wrong, still captures the dominant linear effects. See Yang et al. (2016) and Yang and Zhang (2023) for theoretical analyses.

Lesson 8: The relative advantage of matching over IPW reverses with model misspecification.

	Part 1: Correct PS model		**Part 2: Misspecified PS model**	
Estimator	Bias	RMSE	Bias	RMSE
HJ-ATT	+0.002	0.101	+0.154	0.174
NNM	+0.006	0.121	+0.020	0.139

IPW Needs a Correct PS Model; Matching Needs Covariate Balance

IPW is a *model-dependent* estimator: its consistency relies on the PS model being correctly specified. Matching is a *design-dependent* estimator: its consistency relies on matched pairs being approximately balanced on the true confounders. This condition can hold even when the PS model is wrong, provided the estimated score still separates units well enough. Neither dominates unconditionally. The recommendation is to assess PS model plausibility via balance diagnostics, or to use both as a sensitivity check (Yang et al. 2016; Yang and Zhang 2023).

6.8 The Limits of Propensity Score Methods

6.8.1 The Untestable Assumption

Every result in this chapter rests on unconfoundedness $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$: all confounders of the T–Y relationship are observed and included in X. In graphical terms: X blocks every back-door path from T to Y.

Design-Based vs. Model-Based Identification

Randomization *guarantees* ignorability: by construction, no back-door paths exist. Propensity scores *assume* ignorability: the analyst hopes all confounders have been measured and included in X, but this can never be verified from data alone. This distinction — between identification secured by the study design and identification secured by a modeling assumption — is one of the most important dividing lines in causal inference.

Two complementary responses. *Sensitivity analysis* (Chapter 9) keeps the back-door framework but quantifies how strong an unmeasured confounder would have to be to overturn the conclusion. The *instrumental variable* approach (Chapter 7) abandons the unconfoundedness assumption entirely by exploiting an external source of variation in treatment independent of the unobserved confounders.

6.8.2 The Road to Instrumental Variables

Strategy	Key assumption	Identification mechanism
Back-door adjustment (PS)	All confounders are observed	Condition on X to block $T \leftarrow U \rightarrow Y$
Instrumental variables	Some confounders may be unobserved	Exploit exogenous variation $Z \rightarrow T$

Under IV assumptions, the ratio of the Z-induced change in Y to the Z-induced change in T — the Wald estimator — identifies a causal parameter, typically the LATE among compliers. This is derived formally in Chapter 7.

6.9 Chapter Summary

Symbol	Meaning
$\pi(X)$	Propensity score $P(T{=}1 \mid X)$
$b(X)$	Generic balancing score; $\pi(X)$ is the coarsest
$\widehat{\tau}_{\text{HT}}$	Horvitz–Thompson IPW estimator Equation 6.2
$\widehat{\tau}_{\text{HJ}}$	Hájek (self-normalized) IPW estimator Equation 6.3
NNM	Nearest-neighbor matching estimator

1. **The propensity score reduces dimension.** Under unconfoundedness, $\pi(X)$ is a balancing score ($T \perp\!\!\!\perp X \mid \pi(X)$). By **?@thm-ps**, $(Y(0), Y(1)) \perp\!\!\!\perp T \mid \pi(X)$, so identification requires adjustment for the scalar $\pi(X)$ alone.
2. **IPW identification.** The ATE is identified as $\mathbb{E}[TY/\pi(X)] - \mathbb{E}[(1-T)Y/(1-\pi(X))]$ under strong ignorability. The HT estimator is consistent but sensitive to extreme weights; the Hájek variant is recommended in practice.
3. **Matching targets the ATT.** Nearest-neighbor matching finds control units with similar propensity scores to each treated unit. It provides transparent covariate balance diagnostics but differs from IPW in estimand and methodology.
4. **Positivity is necessary; strong overlap is needed for stable estimation.** Positivity ($0 < \pi(X) < 1$ a.s.) is necessary for ATE identification. Strong overlap ($c \leq \pi(X) \leq 1-c$) ensures stable $\sqrt{n}$-inference. Trimming or re-targeting addresses near-violations.
5. **The fundamental limitation.** Unconfoundedness is untestable. When unobserved confounders are present, either sensitivity analysis or an instrumental variable is needed.

Design	Key assumption	Identified estimand
Randomized experiment	$(Y(0), Y(1)) \perp\!\!\!\perp T$ (by design)	ATE
Propensity-score adjustment	$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$	ATE or ATT
Instrumental variables	Relevance, exogeneity, exclusion	LATE (compliers)

6.10 Problems

1. Propensity score and balancing. Suppose $X = (X_1, X_2)$ with $\text{logit}(\pi(X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

(a) State and prove the Balancing Property $T \perp\!\!\!\perp X \mid \pi(X)$.
(b) Two units i and j have $X_i = (1, 2.3)$ and $X_j = (0, 3.5)$ but $\widehat{\pi}(X_i) = \widehat{\pi}(X_j) = 0.4$. If i is treated and j is control, explain why comparing Y_i and Y_j is a valid approximation to the counterfactual comparison, and what assumption is required.
(c) Explain why matching on $\widehat{\pi}(X)$ is *not* the same as matching on X directly. Under what conditions do they give the same answer?

2. ATE, ATT, and propensity score weighting. A dataset has $n = 1000$ observations with $n_1 = 400$ treated units.

(a) Write the Horvitz–Thompson IPW estimator of the ATE as a weighted sum of observed outcomes Y_i. What weights do treated units receive? What weights do control units receive?
(b) Derive an analogous IPW estimator for the ATT. (*Hint*: the ATT averages over the treated distribution; the weight for control units should reflect the treatment odds.)
(c) Show that when $\pi(X) = p$ for all units, the IPW estimator of the ATE reduces to the difference-in-means estimator. Under what experimental design does $\pi(X) = 0.5$ exactly?

3. Overlap. Let $\pi(X)$ be the true propensity score and suppose $\pi(x_0) = 1$ for some x_0.

(a) For each of τ_{ATE} and τ_{ATT}, identify the formula that fails to be identified when $\pi(x_0) = 1$, and explain why.
(b) Suppose overlap fails only on a set $\mathcal{S}$ with $P(X \in \mathcal{S}) = 0.15$. Define the trimmed ATE estimand. How does it differ from the ATE?
(c) A researcher applies the Crump et al. (2009) rule with $\eta = 0.1$, removing 8% of the sample. List two reasons the trimmed estimator may have lower variance, and state the cost in terms of external validity.

4. Hidden confounding (preview of Chapter 9). Suppose you estimate $\hat{\tau}_{\text{ATE}} = 3.2$ using propensity-score methods, and a referee questions whether unconfoundedness holds.

(a) Define what it means for U to be a "hidden confounder" in the context of the DAG, and explain why its presence invalidates the IPW identification formula.
(b) Explain in words what it means for an estimated effect to be "robust to hidden confounding," and why such an assessment depends on a quantitative yardstick. (Chapter 9 develops three formal yardsticks: Rosenbaum's Γ, the E-value, and the marginal sensitivity model.)
(c) Why does the IV strategy of Chapter 7 avoid the hidden-confounding problem entirely, and what assumption replaces unconfoundedness?

5. Matching vs. IPW. You have $n = 500$ observations comparing Hájek IPW targeting the ATE (estimator A) and 1:1 nearest-neighbor matching targeting the ATT (estimator B).

(a) Explain in one sentence why estimators A and B target different estimands even though both use $\hat{\pi}(X)$.
(b) After matching, the SMD for covariate X_1 is 0.05 (vs. 0.45 before matching). What does this tell you about the success of matching, and what assumption does balance on observed covariates not verify?
(c) Under what condition on treatment effect heterogeneity are the ATE and ATT equal?

Chapter 7

Instrumental Variables

Learning Objectives

By the end of this chapter, students should be able to:

1. Diagnose the endogeneity problem: explain why an unobserved confounder makes back-door adjustment fail, and describe how an instrumental variable provides an alternative identification route.
2. State the three IV assumptions — relevance, exogeneity, and exclusion — in the DAG/do-calculus, structural-equation, and potential-outcomes languages, and explain which are testable and which require institutional justification.
3. Follow how the IV assumptions identify the Wald estimand, and interpret the reduced form and first stage as its two observable components.
4. Explain what the Wald estimand identifies when treatment effects are heterogeneous — the Local Average Treatment Effect for compliers — why that estimand depends on the choice of instrument, and when it coincides with the ATE.
5. Compare IV and back-door adjustment on the assumptions each requires, the estimand each identifies, and the ways each can fail.

7.1 Why Instrumental Variables?

Chapter 6 showed how causal effects can be identified when all confounders are observed and can be blocked by conditioning on X. When some confounders are unobserved, back-door adjustment fails. This chapter develops instrumental variables (IV) as an alternative identification strategy: rather than blocking the confounding path $T \leftarrow U \rightarrow Y$, IV exploits an external variable Z whose effect on T is free of confounding by U. This chapter asks what IV identifies and under what assumptions; Chapter 13 asks how that estimand is computed and tested in practice.

7.1.1 The Endogeneity Problem

Unconfoundedness $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$ requires that every variable affecting both treatment and outcome is observed and included in X. In many empirical settings this is implausible: in labor economics, unobserved ability or motivation affects both schooling decisions and wages; in epidemiology, unobserved health behaviors affect both treatment uptake and outcomes. Whenever an unobserved confounder U creates a back-door path $T \leftarrow U \rightarrow Y$, the adjustment formula fails:

$$\int f(y \mid t, x)\, p(x)\, dx \;\neq\; f(y \mid \mathrm{do}(T{=}t)).$$

The gap is the *endogeneity bias*. We need a different identification strategy.

7.1.2 The IV Idea

An *instrumental variable* Z is an observed variable that: (1) *moves* treatment T (relevance); (2) *does so exogenously* — Z is unrelated to the unobserved confounder U (exogeneity); (3) *affects* Y *only through* T — Z has no direct path to Y (exclusion). Under these three conditions, the variation in T induced by Z is free of confounding by U, so the ratio of the Z-induced variation in Y to the Z-induced variation in T becomes a meaningful target for identification.

> Example: Returns to Schooling
>
> A researcher wants to estimate the causal effect of years of schooling T on log wages Y. Unobserved ability U raises both schooling and wages, so OLS is biased upward. No set of observed covariates X fully captures ability. An instrumental variable Z that shifts schooling for reasons unrelated to ability — such as proximity to a school or a policy change in compulsory attendance laws — provides a way to isolate the causal effect of schooling on wages.

IV does not eliminate the confounding path $T \leftarrow U \rightarrow Y$, nor does it make treatment as-if randomly assigned for the full population. Instead, IV *avoids* confounding rather than controlling for it — a distinction that matters both for interpreting what is identified (the LATE for compliers, not the ATE) and for understanding which assumption does the heaviest lifting (exclusion, not unconfoundedness).

7.2 Graphical Setup and Core Assumptions

7.2.1 The IV DAG

The causal structure for a basic IV model with covariates X is:

```
\usetikzlibrary{arrows.meta, positioning}
\definecolor{accent}{RGB}{46,117,182}
\definecolor{defbg}{RGB}{238,244,251}
\definecolor{backdoorred}{RGB}{192,0,0}
\definecolor{causalgreen}{RGB}{26,122,58}
\definecolor{warnbg}{RGB}{255,240,240}
\tikzset{
  w7obs/.style={circle,draw=accent,line width=1pt,minimum size=10mm,font=\small,fill=defbg},
  w7unobs/.style={circle,draw=backdoorred,line width=1pt,dashed,minimum size=10mm,font=\small,fill=wa
  w7arr/.style={-{Stealth[length=5pt]},line width=1pt,color=accent},
  w7redarr/.style={-{Stealth[length=5pt]},line width=1pt,color=backdoorred,dashed},
  w7grnarr/.style={-{Stealth[length=5pt]},line width=1pt,color=causalgreen}
}
\begin{tikzpicture}[node distance=14mm and 20mm]
  \node[w7obs] (Z) {$Z$};
  \node[w7obs,right=of Z] (T) {$T$};
  \node[w7obs,right=of T] (Y) {$Y$};
  \node[w7obs,below=12mm of T] (X) {$X$};
  \node[w7unobs,above right=10mm and 10mm of T] (U) {$U$};
  \draw[w7arr] (Z) -- node[above,font=\scriptsize\color{accent}]{(1)} (T);
  \draw[w7arr] (T) -- (Y);
  \draw[w7grnarr] (X) -- (T);
  \draw[w7grnarr] (X) -- (Y);
  \draw[w7grnarr] (X) -- (Z);
  \draw[w7redarr] (U) -- (T);
  \draw[w7redarr] (U) -- (Y);
\end{tikzpicture}
```

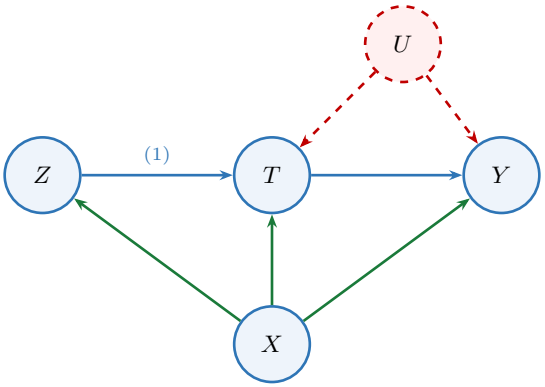


Figure 7.1: The basic IV DAG. Blue: causal effects; dashed red: confounding paths; green: covariate effects. The label (1) marks relevance; the absence of $Z \to Y$ encodes exclusion.

The three IV assumptions correspond to three distinct features of this DAG: (1) **Relevance**: there is a directed path $Z \to T$; (2) **Exogeneity**: conditional on X, the DAG implies $Z \perp\!\!\!\perp U \mid X$ by d-separation; (3) **Exclusion**: every directed path from Z to Y in $\mathcal{G}$ passes through T — there is no direct edge $Z \to Y$.

The distinction between exogeneity and exclusion is fundamental. Exogeneity says the instrument is not confounded with latent causes of the outcome. Exclusion says the instrument has no causal channel to the outcome except through treatment. A randomized encouragement may be exogenous by design, yet still violate exclusion if the encouragement changes outcomes through information, motivation, or stigma apart from the treatment itself.

7.2.2 The Three Assumptions in Three Languages

The same three assumptions can be expressed in three causal languages. These are parallel formulations, not literally identical statements: each highlights a different aspect of the design. The graphical formulation is most useful for causal design; the structural formulation is most useful for deriving moment restrictions; the potential-outcomes formulation prepares the ground for the LATE framework. The assumptions are ordered *relevance* → *exogeneity* → *exclusion*, reflecting the natural sequence in which a researcher assesses them.

Assumption	**Potential outcomes**	**Structural / econometric**	**Do-calculus / DAG**
Relevance	$P(T_i(1) \neq T_i(0)) > 0$	$\pi \neq 0$ in $T = \pi Z + \delta^\top X + \eta$	$Z \to T$ in $\mathcal{G}$ (no d-separation)
Exogeneity	$Z \perp\!\!\!\perp (Y(0), Y(1), T(0), T(1)) \mid X$	$\mathbb{E}[\varepsilon \mid Z, X] = 0$	$Z \perp\!\!\!\perp U \mid X$
Exclusion	$Y_i(t, z) = Y_i(t, z')$ for all z, z'	Z absent from structural equation for Y	$f(y \mid x, \mathrm{do}(T{=}t), z) = f(y \mid x, \mathrm{do}(T{=}t))$

Remark: Three Languages, One Identification Argument

These three formulations are aligned but not literally identical. The graphical statement encodes causal structure. The potential-outcomes statement encodes counterfactual independence. The structural formulation encodes moment orthogonality. In a well-specified IV model they support the same identification argument, but they are not interchangeable symbols.
Relevance is about the $Z \to T$ link — Z must genuinely move T. *Exogeneity* is about the absence of omitted common causes linking Z to Y. *Exclusion* is about the absence of any direct causal path $Z \to Y$ that bypasses T.

Why the do-calculus formulation is preferred. The exclusion restriction in the do-calculus column reads:

$$f(y \mid x, \mathrm{do}(T{=}t), z) = f(y \mid x, \mathrm{do}(T{=}t)).$$

This is a statement about the *interventional* density — the distribution of Y after we have *set* $T = t$ by do-surgery. The do-operator makes it impossible to confuse this with the observational statement $f(y \mid x, T{=}t, z) = f(y \mid x, T{=}t)$, which is a much weaker condition.

7.2.3 Relevance

Definition: Relevance

The instrument Z is **relevant** if it has a non-zero causal effect on the treatment T within at least some stratum of covariates X: $P(T \mid \mathrm{do}(Z{=}z), X{=}x)$ varies with z for some x. In the linear first-stage model $T = \pi Z + \delta^\top X + \eta$, this reduces to $\pi \neq 0$.

Graphically, relevance means Z and T are not d-separated in $\mathcal{G}$. The conditional covariance $\mathrm{Cov}(Z, T \mid X)$ and the first-stage F-statistic are empirical diagnostics for the observable association; the causal claim that Z shifts T is design-based.

7.2.4 Exogeneity

Definition: Exogeneity

The instrument Z is **exogenous** given covariates X if there is no open back-door path from Z to Y through unobserved causes: $Z \perp\!\!\!\perp U \mid X$ in $\mathcal{G}$ by d-separation.

In the structural linear model, the same requirement appears as $\mathbb{E}[Z\varepsilon \mid X] = 0$, because the structural residual ε is a function of U. This moment condition is a *consequence* of the graphical assumption, not a definition of exogeneity. A researcher who adopts the moment condition as a primitive has no guarantee that Z is free of back-door paths to the outcome.

Example: Quarter of Birth [@angrist1991does]

Angrist and Krueger (1991) used quarter of birth as an instrument for schooling to estimate the return to education. Exogeneity requires no open back-door path from quarter of birth to wages through unobserved variables such as ability or family background; birth timing is largely outside parental control. Bound et al. (1995) later raised concerns about weak first stages in some specifications.

Exogeneity is untestable: the unobserved confounder U is, by definition, unobserved.

7.2.5 Exclusion

Definition: Exclusion Restriction

The instrument Z satisfies the **exclusion restriction** if it affects the outcome Y *only* through its effect on the treatment T. In do-calculus terms:

$$f(y \mid x, \mathrm{do}(T{=}t), z) = f(y \mid x, \mathrm{do}(T{=}t)) \quad \text{for all } z.$$

In potential outcomes terms: $Y_i(t, z) = Y_i(t, z')$ for all $z \neq z'$.

Exclusion is a *causal* restriction, not a conditional-correlation restriction. Adding a direct arrow $Z \to Y$ to the DAG is exactly the formal counterpart of exclusion failing. In the mutilated graph $\mathcal{G}_{\overline{T}}$, the path $Z \to Y$ would remain open. Rule 1 of the do-calculus, which would allow removing Z from the density of Y given $\mathrm{do}(T{=}t)$, no longer applies.

Why Exclusion Is the Hardest Assumption

Relevance can be tested. Exogeneity can sometimes be defended by design. But exclusion — that Z has *no direct effect* on Y whatsoever — is:

- **Untestable** in just-identified models.
- **Easy to violate in practice.** An instrument that "moves treatment" often also moves other inputs. If Z is distance to a hospital (instrument for treatment uptake), it may also directly affect health outcomes through travel time in emergencies.
- **Consequential.** Even a small direct effect of Z on Y can cause substantial bias in the Wald estimand, especially when the first stage is weak (derived in Section 7.4).

Remark: Three Assumptions Are Necessary but Not Sufficient

The three IV assumptions are necessary but not sufficient for *point* identification of any causal estimand (Hernán and Robins 2006; Levis et al. 2024). A fourth *structural assumption* on the counterfactual distribution of treatment response or treatment effect heterogeneity is required. The choice of fourth assumption determines which causal quantity the Wald estimand identifies: constant treatment effects (Section 8.7), under which the Wald estimand identifies the ATE; or monotonicity (Section 7.7), under which it identifies the LATE for compliers.
Strikingly, across these different identification schemes the same conditional Wald formula $\text{Cov}(Z, Y \mid X)/\text{Cov}(Z, T \mid X)$ appears as the identifying expression in many cases — but its causal target changes. The three core IV assumptions are all expressible within the graphical language; the fourth assumption, by contrast, lies *outside* the graphical language in every case.

Example: Same Observables, Different ATEs

Let $Z, T \in \{0, 1\}$ with $\mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0] = 0.2$ and $\mathbb{E}[Y \mid Z{=}1] - \mathbb{E}[Y \mid Z{=}0] = 0.1$, giving Wald ratio $= 0.5$. Both DGPs share the same compliance structure: $P(\text{co}) = 0.2$, $P(\text{at}) = 0.4$, $P(\text{nt}) = 0.4$.
DGP A (constant effects): $\tau_i = 0.5$ for all units. ATE $= 0.5$. Wald ratio $= 0.5$.
DGP B (heterogeneous effects + monotonicity): $\tau_{\text{co}} = 0.5$, $\tau_{\text{at}} = \tau_{\text{nt}} = 0$. ATE $= 0.2 \times 0.5 = 0.1$. Wald ratio $= 0.5$.
Both DGPs produce the same first stage, reduced form, and Wald ratio. Yet the ATE is 0.5 under DGP A and 0.1 under DGP B — a fivefold difference. No amount of data can distinguish them without a fourth structural assumption.

7.3 Identification in the Linear Homogeneous-Effect Model

Framework 1: Constant Treatment Effect and Linear Structure

This section works within a linear structural model in which the treatment effect is the *same* for every unit: $Y_i(t) - Y_i(t') = \beta(t - t')$ for all i, t, t'. Under this assumption, IV identifies the single parameter β, which equals the ATE, the ATT, and the LATE simultaneously: $\beta = \text{ATE} = \text{ATT} = \text{LATE}$.

7.3.1 The Linear Structural Model

Consider the linear structural model:

$$Y = \alpha + \beta T + \gamma^\top X + \varepsilon, \tag{7.1}$$

$$T = \pi Z + \delta^\top X + \eta, \tag{7.2}$$

where ε and η are structural errors with $\text{Cov}(\varepsilon, \eta) = \rho\sigma\tau \neq 0$. The non-zero covariance is the source of endogeneity: OLS applied to Equation 7.1 gives a biased estimator of β. The *reduced form* substitutes Equation 13.2 into Equation 7.1:

$$Y = \alpha + \beta\pi Z + (\beta\delta + \gamma)^\top X + (\beta\eta + \varepsilon).$$

The reduced-form coefficient on Z is $\beta\pi$: the total effect of the instrument on the outcome. Dividing by the first-stage coefficient π recovers β, provided $\pi \neq 0$.

7.3.2 The OLS Bias

The probability limit of the OLS estimator is:

$$\operatorname{plim} \hat{\beta}_{\text{OLS}} = \beta + \frac{\operatorname{Cov}(T, \varepsilon)}{\operatorname{Var}(T)} = \beta + \frac{\rho\sigma\tau}{\pi^2 \operatorname{Var}(Z) + \tau^2}.$$

The bias is zero only if $\rho = 0$ (no unobserved confounding) or if the instrument perfectly determines T. The direction of the bias depends on the sign of ρ.

7.3.3 Derivation of the Wald Estimand

The derivation has three ingredients: the first stage (how much Z moves T), the reduced form (how much Z moves Y), and the exclusion restriction (any effect of Z on Y must operate through T).

1. **Exogeneity** ($Z \perp\!\!\!\perp U \mid X$) implies $\mathbb{E}[\varepsilon \mid Z, X] = 0$.
2. **Exclusion** (no Z term in the outcome equation) combined with exogeneity yields the moment condition $\mathbb{E}[\varepsilon \cdot Z \mid X] = 0$.
3. Multiplying Equation 7.1 by $(Z - \mathbb{E}[Z \mid X])$ and applying step 2: $\operatorname{Cov}(Y, Z \mid X) = \beta \cdot \operatorname{Cov}(T, Z \mid X)$, which is the **reduced-form decomposition**: the Z–Y covariance is entirely attributable to the causal path $Z \to T \to Y$.
4. **Relevance** ($\operatorname{Cov}(Z, T \mid X) \neq 0$) ensures the denominator is non-zero:

$$\beta = \frac{\operatorname{Cov}(Y,\, Z \mid X)}{\operatorname{Cov}(T,\, Z \mid X)}. \tag{7.3}$$

In the binary instrument case ($Z \in \{0, 1\}$) without covariates, this simplifies to the **Wald estimand**:

The Wald Estimand

$$\beta = \frac{\mathbb{E}[Y \mid Z{=}1] - \mathbb{E}[Y \mid Z{=}0]}{\mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0]}. \tag{7.4}$$

The numerator is the *reduced form*: the total effect of Z on Y. The denominator is the *first stage*: the effect of Z on T. The exclusion restriction guarantees the entire reduced form operates through T; dividing by the first stage strips out the $Z \to T$ piece.

Remark: Estimation Deferred

The Wald estimand is an identification result: it expresses β as a ratio of observable quantities. Estimation — how to consistently estimate this ratio from a finite sample, including the correct treatment of standard errors — is taken up in Chapter 13.

Example: Returns to Schooling [@angrist1991does]

In the Mincer earnings equation $\log W = \alpha + \beta S + \gamma^\top X + \varepsilon$, where S is years of schooling and W is wages, OLS is biased upward because unobserved ability U raises both S and W. Angrist and Krueger (1991) use quarter of birth as Z: proximity to mandatory school-leaving age at different birth quarters generates exogenous variation in completed schooling. The IV estimate of the return to schooling is approximately 0.08–0.10 per year.

7.4 Why the IV Assumptions Matter

Each assumption is load-bearing, and each failure mode produces a distinct, quantifiable distortion of the Wald estimand.

When relevance fails. If $\pi = 0$, the Wald estimand is undefined. When π is small but nonzero, the instrument is *weak*. The estimator's variance diverges as $\pi \to 0$, and finite-sample bias pulls the IV estimate toward the OLS estimate at a rate proportional to $1/F$, where F is the first-stage F-statistic.

When exclusion fails. If $Y = \alpha + \beta T + \delta Z + \gamma^\top X + \varepsilon$ with $\delta \neq 0$, the Wald estimand converges to $\beta + \delta/\pi$. The bias δ/π is amplified by a weak first stage: a small direct effect combined with a weak instrument can produce large bias. This is why a weak instrument with a plausible exclusion violation is not "nearly valid" — it may be severely misleading.

When exogeneity fails. If $\text{Cov}(Z, \varepsilon \mid X) \neq 0$, the Wald estimand converges to $\beta + \text{Cov}(\varepsilon, Z)/\text{Cov}(T, Z)$. Again the bias is amplified by weak instruments.

Assumption	Directly testable?	Basis for assessment
Relevance	Association testable; causal claim design-based	First-stage F-statistic; the causal $Z \to T$ link rests on the design
Exogeneity	No	Institutional knowledge; randomization (if available); placebo regressions on pre-determined outcomes
Exclusion	No in just-identified case; partially in overidentified case	Institutional argument; overidentification test (J-test, Chapter 13) checks mutual consistency but cannot confirm all instruments are valid (Kitagawa 2015)

7.5 Lab: OLS vs. IV Across Instrument Strengths

This lab verifies the bias formulas numerically and traces the bias–variance tradeoff across the full range of instrument strength. It studies estimator behavior *conditional* on the IV assumptions being true; it does not address whether a proposed instrument is valid in an applied study.

DGP for Lab 7

The simulation implements the linear structural model of Section 8.7 with no covariates X. Each replication draws $n = 500$ observations from:

$$Y_i = \beta T_i + \varepsilon_i, \qquad T_i = \pi Z_i + \eta_i,$$

with $Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ and structural errors:

$$\eta_i = U_i, \qquad \varepsilon_i = \rho U_i + \sqrt{1-\rho^2}\,\xi_i, \qquad U_i, \xi_i \overset{\text{i.i.d.}}{\sim} N(0,1).$$

Fixed parameters: $\beta = 1$ (true causal effect), $\rho = 0.8$ (strong positive endogeneity). The first-stage coefficient π is varied across eight values $\pi \in \{0, 0.10, 0.15, 0.20, 0.30, 0.50, 1.00, 2.00\}$. The expected first-stage F-statistic is approximately $n\pi^2 = 500\pi^2$.

Potential Outcomes Interpretation

The potential outcome is $Y_i(t) = \beta t + \varepsilon_i$, so the individual treatment effect is constant: $Y_i(t) - Y_i(t') = \beta(t - t')$ for all i. This is Framework 1: β is simultaneously the ATE, ATT, and LATE.
The shared factor U_i creates confounding: $\text{Cov}(Y_i(t), T_i) = \text{Cov}(\varepsilon_i, \eta_i) = \rho \neq 0$, so unconfoundedness fails. The instrument Z_i is drawn independently of U_i, so $Z_i \perp\!\!\!\perp \varepsilon_i$ (exogeneity); Z_i does not appear in $Y_i(t)$ (exclusion); and Z_i moves T_i through π (relevance). All three IV assumptions hold exactly by construction.

Estimators. OLS regresses Y on T: $\hat{\beta}_{\text{OLS}} = \text{Cov}(Y, T)/\text{Var}(T)$, converging to $1 + 0.8/(\pi^2 + 1)$. IV uses the Wald estimator: $\hat{\beta}_{\text{IV}} = \text{Cov}(Y, Z)/\text{Cov}(T, Z)$, consistent for β for any $\pi \neq 0$.

Results ($n = 500$, $B = 2{,}000$ replications, seed 2024):

π	F	Theory bias	OLS mean	OLS RMSE	IV mean	IV RMSE
0.00	1	+0.800	1.801	0.802	—	—
0.10	6	+0.792	1.791	0.792	0.872	4.393
0.15	13	+0.782	1.782	0.783	0.764	6.491
0.20	21	+0.769	1.769	0.770	0.940	0.385
0.30	46	+0.734	1.735	0.735	0.975	0.165
0.50	127	+0.640	1.639	0.640	0.996	0.091
1.00	500	+0.400	1.401	0.402	0.999	0.045
2.00	2006	+0.160	1.160	0.161	1.000	0.022

Lesson 1: The OLS bias formula is exact. The theory bias $0.8/(\pi^2 + 1)$ matches the simulated OLS bias to four decimal places across all eight values of π. OLS is biased in the direction of ρ at every value of π, including $\pi = 0$.

Lesson 2: IV is consistent but has catastrophically heavy tails when the instrument is weak. At $\pi = 0.10$ ($F \approx 6$) and $\pi = 0.15$ ($F \approx 13$), the IV *mean* is far from the true value. The IV *median* is ≈ 1.01 at both values, confirming consistency — the mean is dragged off by a small fraction of replications in which the first stage is near zero. Standard deviations of 4.4 and 6.5 make these estimates useless.

Lesson 3: The RMSE crossover occurs near $F \approx 20$. IV first beats OLS on RMSE at $\pi = 0.20$ ($F \approx 21$): $0.385 < 0.770$. The Staiger–Stock rule of thumb ($F \geq 10$) is slightly too lenient: at $F \approx 13$, IV RMSE is still 6.5, nearly 8× OLS. A more conservative threshold of $F \geq 20$–25 is needed.

Lesson 4: A strong instrument eliminates both problems. At $\pi = 1.00$ ($F \approx 500$), IV RMSE $= 0.045$ while OLS RMSE $= 0.402$ — a **9-fold** improvement from IV. OLS efficiency is illusory: its small variance is offset by a large, persistent bias.

RMSE Cannot Be the Only Criterion

OLS has lower RMSE than IV for $\pi < 0.20$ in this DGP. This does not vindicate OLS. An estimator with RMSE $= 0.78$ because it is biased by 0.80 is useless for causal inference: the bias is systematic and does not shrink with sample size. As $n \to \infty$, IV RMSE shrinks to zero while OLS RMSE stays at 0.80. The RMSE comparison is only meaningful in finite samples — it quantifies when a weak instrument is so unreliable that IV is not yet practically useful, but that is an argument for finding a stronger instrument, not for using OLS.

Remark: The First-Stage F-Statistic as a Diagnostic

The formula $F \approx n\pi^2$ gives first-stage F-statistics matching the simulation throughout. The Bound et al. (1995) threshold of $F \geq 10$ is widely used in practice and corresponds roughly to IV RMSE within a factor of two of the strong-instrument limit. This simulation suggests that threshold may understate the problem when endogeneity is strong ($\rho = 0.8$): at $F = 13$ the IV RMSE is 6.5, far above any useful threshold. A researcher should report the first-stage F alongside any IV estimate, and treat values below 20–25 with particular caution.

7.6 Multiple Instruments and Overidentification

When there are exactly as many instruments as endogenous variables ($q = p$), the model is *just-identified*. When $q > p$, the model is *overidentified*: the extra instruments impose additional moment restrictions that are testable. Under the homogeneous-effect linear model, every valid instrument must imply the same structural coefficient β, so those extra restrictions are testable. Under heterogeneous treatment effects, valid instruments can legitimately identify *different* LATEs because they shift treatment for different complier populations.

The **order condition** ($q \geq p$, counting requirement) and the **rank condition** (instruments are linearly independent in the first stage) are both necessary for identification. The Sargan–Hansen J-test (Sargan

1958; Hansen 1982) formalizes the mutual-consistency check; rejection indicates at least one instrument either violates exogeneity or exclusion, or identifies a different LATE, but does not localize which. The test statistic and asymptotic distribution are derived in Chapter 13.

What Overidentification Does Not Do

Passing the J-test does not confirm that any instrument is valid. It only shows that the sample moment conditions are mutually compatible — a weaker conclusion than validity. Multiple invalid instruments can agree with one another if they share the same violation; consistent instruments are not necessarily valid ones. Overidentification is an opportunity for a consistency check, not a substitute for the institutional argument that makes a design credible.

7.7 Heterogeneous Treatment Effects and the LATE Framework

Framework 2: Heterogeneous Treatment Effects

We now drop homogeneity and allow $\tau_i = Y_i(1) - Y_i(0)$ to vary across units. Throughout this section we work in the binary instrument, binary treatment setting ($Z, T \in \{0, 1\}$). In this setting, the Wald ratio generally no longer identifies the ATE. Instead, under an additional assumption of monotonicity, it identifies the *Local Average Treatment Effect* (LATE): the average treatment effect for the subpopulation whose treatment status is changed by the instrument.

7.7.1 Compliance Types

Definition: Compliance Types [@angrist1996identification]

For binary $Z, T \in \{0, 1\}$, the four compliance types are defined by the pair of potential treatment decisions $(T_i(0), T_i(1))$:

Compliance type	$T_i(0)$	$T_i(1)$
Complier	0	1
Always-taker	1	1
Never-taker	0	0
Defier	1	0

A **complier** takes treatment if and only if the instrument is switched on. An **always-taker** takes treatment regardless of Z. A **never-taker** never takes treatment. A **defier** does the opposite of what the instrument suggests.

The instrument Z only shifts treatment for *compliers*: always-takers and never-takers have the same treatment status regardless of Z, so they contribute nothing to the denominator $\mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0]$.

7.7.2 The Monotonicity Assumption

Definition: Monotonicity [@imbens1994identification]

The treatment assignment is **monotone in** Z if $T_i(1) \geq T_i(0)$ for all i. Equivalently: there are no defiers in the population.

Monotonicity is not a generic law of causal inference. It is a design-specific claim about how this particular instrument changes treatment behavior. Switching the instrument from 0 to 1 may induce some units to take treatment (compliers) and leave others unaffected (always-takers or never-takers), but it should not reverse anyone's treatment decision.

7.7.3 The LATE Theorem

Theorem: LATE Theorem [@imbens1994identification]

Suppose: (1) *Exogeneity (PO form):* $Z \perp\!\!\!\perp (Y(0), Y(1), T(0), T(1))$; (2) *Exclusion:* $Y_i(t,z) = Y_i(t)$ for all z; (3) *Relevance:* $\mathbb{E}[T(1) - T(0)] \neq 0$; (4) *Monotonicity:* $T_i(1) \geq T_i(0)$ for all i. Then the Wald estimand identifies the **Local Average Treatment Effect** (LATE):

$$\frac{\mathbb{E}[Y \mid Z{=}1] - \mathbb{E}[Y \mid Z{=}0]}{\mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0]} = \mathbb{E}[Y(1) - Y(0) \mid T_i(1) > T_i(0)] \equiv \tau_{\text{LATE}}.$$

Proof

Denominator. By consistency for T and exogeneity ($Z \perp\!\!\!\perp (T(0), T(1))$):

$$\mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0] = \mathbb{E}[T(1)] - \mathbb{E}[T(0)] = \mathbb{E}[T(1) - T(0)].$$

Monotonicity (no defiers) gives $\mathbb{E}[T(1){-}T(0)] = P(\text{complier})$, since always-takers contribute $1{-}1 = 0$ and never-takers contribute $0 - 0 = 0$: $\mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0] = P(\text{complier})$.
Numerator. By consistency, exclusion, and exogeneity:

$$\mathbb{E}[Y \mid Z{=}1] - \mathbb{E}[Y \mid Z{=}0] = \mathbb{E}[Y(T(1))] - \mathbb{E}[Y(T(0))] = \sum_c P(c)\, \mathbb{E}[Y(T_c(1)) - Y(T_c(0)) \mid \text{type} = c].$$

For always-takers: $T(1) = T(0) = 1$, contribution $= 0$. For never-takers: $T(1) = T(0) = 0$, contribution $= 0$. For compliers: $T(1) = 1$, $T(0) = 0$, so $Y(T(1)) - Y(T(0)) = Y(1) - Y(0)$. No defiers by monotonicity. Therefore:

$$\mathbb{E}[Y \mid Z{=}1] - \mathbb{E}[Y \mid Z{=}0] = P(\text{complier})\, \mathbb{E}[Y(1) - Y(0) \mid \text{complier}].$$

Ratio. Dividing numerator by denominator: $\tau_{\text{LATE}} = \mathbb{E}[Y(1) - Y(0) \mid \text{complier}]$. □

Remark: The Two Senses of "Local"

The LATE is local in two senses: it is local to the complier group defined by *this* instrument, and local to the *particular instrument* that defines that group. A different instrument, even for the same treatment, will generally select a different complier population and identify a different LATE. This is developed in Section 7.8.

7.8 Interpreting IV Estimands

7.8.1 What the Two Frameworks Say

	Framework 1 (linear, homogeneous)	**Framework 2** (heterogeneous effects)
Key assumption	$\tau_i = \beta$ for all i	Monotonicity; no defiers
What IV identifies	$\beta = \text{ATE} = \text{ATT} = \text{LATE}$	$\tau_{\text{LATE}} = \mathbb{E}[\tau_i \mid \text{complier}]$
Estimand depends on instrument?	No (same β regardless of Z)	Yes (different $Z \Rightarrow$ different compliers $\Rightarrow$ different LATE)
Required for ATE?	Yes, automatically	Only if all units are compliers or effects homogeneous

Framework 1 is a special case of Framework 2: when $\tau_i = \beta$ for all i, the LATE equals the ATE equals β. In applied work, the default interpretation of the Wald estimand is the LATE; the ATE interpretation requires the additional homogeneity argument of Framework 1.

Remark: Same Formula, Different Estimand

Across a range of fourth assumptions, the conditional Wald formula $\text{Cov}(Z, Y \mid X)/\text{Cov}(Z, T \mid X)$ serves as the identifying expression in many cases (Levis et al. 2024). Two researchers using the same instrument and computing the same Wald ratio may be consistently estimating different causal quantities if they maintain different structural assumptions. The data alone cannot resolve which estimand the Wald ratio identifies; that determination requires the researcher to commit to a structural assumption about treatment response.

7.8.2 When Does LATE Equal ATE?

LATE equals ATE only under additional structure, most notably treatment-effect homogeneity. To see this, decompose the ATE by compliance type:

$$\text{ATE} = P(\text{co})\,\mathbb{E}[\tau_i \mid \text{co}] + P(\text{at})\,\mathbb{E}[\tau_i \mid \text{at}] + P(\text{nt})\,\mathbb{E}[\tau_i \mid \text{nt}].$$

The LATE equals only the first term divided by its probability weight. ATE and LATE coincide if and only if: (1) mean treatment effects are equal across compliance types, or (2) everyone is a complier, or (3) the average effects for always-takers and never-takers happen to equal the LATE — an untestable coincidence.

7.8.3 Different Instruments, Different Estimands

Because the LATE is specific to the complier population, and different instruments select different complier populations, two valid instruments for the same treatment can legitimately identify different LATEs. This is informative about treatment effect heterogeneity, not a contradiction.

Example: Compulsory Schooling Laws vs. Distance to College [@angrist1991does; @card1995using]

Compulsory schooling laws (Angrist and Krueger 1991) identify the return to schooling for students at the margin of dropping out — typically lower-income students. Distance to the nearest college (Card 1995) identifies the return for students deterred by geographic distance — again, disproportionately lower-income. Both instruments are valid; the LATEs can legitimately differ even if both are valid.

7.8.4 The Policy Relevance of LATE

For many policy questions, the LATE is exactly the right estimand. If a policy is designed to encourage a subset of the population to take treatment, then the effect on compliers (those who respond to the encouragement) is precisely what the policy-maker wants to know. When the ATE over the full population is required, IV alone is insufficient under heterogeneous effects; additional assumptions or a second instrument are needed to extrapolate from the LATE to the ATE.

7.9 Practical Guidance on Defending an IV Design

A researcher proposing an instrument should be able to answer five questions explicitly:

1. **What exactly is the instrument?** Specify Z precisely: its source of variation, the level at which it varies, and the population to which it applies.
2. **Why does it shift treatment?** Articulate the causal mechanism by which Z moves T. The first-stage F-statistic is a diagnostic for instrument strength, not a substitute for a causal account of the $Z \to T$ link.
3. **Why is it as-if random relative to latent outcome determinants?** The most credible sources are designed randomization (lotteries, randomized encouragement), natural experiments, and shift-share designs (Bartik 1991; Goldsmith-Pinkham et al. 2020). Placebo regressions on pre-determined outcomes provide partial evidence.
4. **Why can it affect the outcome only through treatment?** Exclusion is untestable in just-identified models. A useful diagnostic: how large would the direct effect δ have to be, relative to π, to overturn the estimated causal effect? When the first stage is weak, the answer is: not very large.

5. **What population margin does it shift?** Identify the complier population. This determines the LATE that is being identified and governs the external validity of the estimates.

7.10 Applied Example: Charter School Lotteries and the KIPP Lynn Study

This example illustrates a canonical randomized-encouragement IV design. The lottery randomizes *offer status* Z, not actual treatment S (years of KIPP attendance). Winning the lottery does not force a student to attend KIPP, and losing does not make later attendance impossible. Thus the lottery offer is the instrument and actual attendance is the endogenous treatment.

Setting. KIPP (Knowledge Is Power Program) schools follow a "No Excuses" model: extended school days, longer academic year, selective teacher hiring, and strict behavioral norms. KIPP Academy Lynn was substantially oversubscribed beginning in 2005. Massachusetts law requires oversubscribed charter schools to select students by lottery, so the school conducted randomized admissions lotteries from 2005 through 2008. The outcome Y_{igt} is the student's standardized score on the Massachusetts MCAS, normalized to mean zero and standard deviation one within each subject–grade–year cell statewide.

Mapping the three assumptions (Angrist et al. 2012).

Relevance. Lottery winners were offered a seat and about 80% accepted; losers rarely enrolled elsewhere at KIPP. The first-stage regression yields a coefficient of approximately 1.2: lottery winners had spent about 1.2 more years at KIPP than lottery losers at the time of each MCAS exam. The first-stage F-statistic is far above conventional thresholds.

Exogeneity. Offer status was determined by randomly drawn lottery-sequence numbers, so independence holds by design. A joint test of covariate balance yields $p = 0.615$, consistent with no pre-lottery differences.

Exclusion. The offer merely provides access to a school; it does not itself deliver instruction. One potential violation is a discouragement effect: losing the lottery might demoralize students. The authors address this by noting that scores of lottery losers are typical of demographically comparable students in Lynn, inconsistent with large discouragement effects. Random assignment of the instrument does not, by itself, imply exclusion — it only guarantees exogeneity.

First stage, reduced form, and 2SLS. The model is just-identified (one excluded instrument per endogenous variable), so the 2SLS estimator of θ (effect per year at KIPP) equals the ratio of the reduced-form coefficient on Z_i to the first-stage coefficient π.

Subject	First stage	Reduced form	2SLS
Math	1.221 (0.068)	0.430 (0.067)	0.352 (0.053)
ELA	1.228 (0.068)	0.164 (0.073)	0.133 (0.059)

Standard errors clustered at the student level. $N = 833$ *student-by-test observations.*

Each year at KIPP raises math scores by approximately 0.35σ and ELA scores by approximately 0.13σ. The reduced-form estimate for math (0.43σ) is larger than the 2SLS estimate (0.35σ) because the first stage exceeds 1: lottery winners accumulated somewhat more than one additional year at KIPP per unit of follow-up time.

LATE interpretation. The 2SLS estimand θ is not the effect of KIPP on all students in Lynn, nor on all applicants — it is the average per-year treatment effect for the *lottery compliers*. Compliance is partial in both directions: some lottery winners do not enroll and some lottery losers eventually find entry.

Treatment effect heterogeneity. Reading gains ($\approx 0.13\sigma$ overall) are driven almost entirely by students classified as having limited English proficiency (LEP, $\approx 0.43\sigma$) and special education needs (SPED, $\approx 0.27\sigma$); non-LEP, non-SPED students show negligible ELA gains. Math effects are large and positive across all subgroups but are largest for LEP and lower-achieving students. This connects directly to Section 7.8: different instruments would identify distinct subpopulation LATEs; the overall 2SLS estimate is a weighted average of subgroup LATEs with weights proportional to each subgroup's share of the complier population.

Remark: Lottery Design and Assumption Credibility [@angrist2009mostly]

Instruments derived from *designed* randomization — lotteries, random assignment, randomized encouragement — provide the most transparent basis for the exogeneity assumption. Exogeneity within the applicant sample, conditional on application cohort, is not merely plausible — it follows from the randomization protocol. This is why lottery-based IV designs occupy a privileged position in the program evaluation literature. The exclusion restriction still requires institutional argument.

7.11 IV versus Back-Door Adjustment

Dimension	Back-door / propensity score	Instrumental variables
Core assumption	All confounders observed: $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$	Valid instrument: relevance, exogeneity, exclusion
Unobserved confounders	Fatal: back-door adjustment fails	Permitted: IV routes around U
Estimand	ATE, ATT, or ATC; all coincide under homogeneity	LATE (compliers only); reduces to common β under homogeneous effects
Testability	Unconfoundedness is untestable; overlap is testable	Relevance testable; exogeneity and exclusion untestable (just-identified)
Main threat	Unmeasured confounder	Exclusion restriction violation
Identifies ATE?	Yes, under strong ignorability	Only under homogeneous effects

Complementary failure modes. Back-door adjustment fails when X does not capture all confounders. IV fails when the exclusion restriction is violated: the bias in the Wald estimand is δ/π, amplified by weak instruments. The two failures are orthogonal: back-door adjustment requires many observed covariates but tolerates no unobserved ones, while IV tolerates unobserved confounders but requires an instrument with no direct effect on the outcome.

When both strategies are available. The Hausman (1978) endogeneity test compares OLS and IV: under the null that T is exogenous given X, both are consistent, and a large discrepancy is evidence of endogeneity. Under the null, an appropriately scaled quadratic form in $\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}$ is asymptotically χ^2_p. Disagreement does not by itself tell us which method is wrong: the two strategies typically target different estimands (ATE vs. LATE), and disagreement can arise from legitimate effect heterogeneity rather than failure of either assumption.

7.12 Chapter Summary

Symbol	Meaning
Z	Instrument
π	First-stage coefficient: $\mathbb{E}[\partial T/\partial Z]$
ρ	Endogeneity: $\text{Cov}(\varepsilon, \eta)/(\sigma\tau)$
τ_{LATE}	$\mathbb{E}[Y(1) - Y(0) \mid T_i(1) > T_i(0)]$
Wald estimand	$(\mathbb{E}[Y \mid Z{=}1] - \mathbb{E}[Y \mid Z{=}0])/(\mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0])$
Reduced form	Total effect of Z on Y
First stage	Effect of Z on T

1. **IV identifies effects from exogenous treatment variation.** When back-door adjustment fails because an unobserved U creates a path $T \leftarrow U \rightarrow Y$, a valid instrument Z identifies the causal effect by exploiting only the component of treatment variation that Z induces. IV does not block the confounding path — it avoids it.

2. **Three assumptions, ordered by testability.** Relevance can be assessed with the first-stage F-statistic (though the F-statistic is a sample diagnostic). Exogeneity and exclusion must be defended by institutional knowledge, design logic, and causal structure. Each violated assumption produces a distinct, quantifiable bias, amplified by weak instruments.
3. **Framework 1: homogeneous-effect SEM $\Rightarrow$ Wald identifies β.** Under constant treatment effects and the linear structural model, the Wald estimand identifies $\beta = \text{ATE} = \text{ATT} = \text{LATE}$.
4. **Framework 2: heterogeneity + monotonicity $\Rightarrow$ Wald identifies LATE.** Under heterogeneous treatment effects and monotonicity, the Wald estimand identifies the average treatment effect for *compliers* only — those whose treatment status changes with the instrument, a latent subgroup defined by $(T(0), T(1))$.
5. **Different instruments identify different effects.** The LATE depends on the instrument through the complier population it selects. This is informative about treatment effect heterogeneity, not a contradiction. LATE equals ATE only under additional structure.
6. **IV versus back-door adjustment.** Complementary failure modes: back-door fails when confounders are unobserved; IV fails when the exclusion restriction is violated or the instrument is not truly exogenous.
7. **Estimation deferred to Chapter 13.** This chapter establishes what IV identifies and under what assumptions. How the Wald ratio is estimated from finite data — reduced-form regression, two-stage least squares, asymptotic inference, and overidentification tests — is the subject of Chapter 13.

7.13 Problems

1. The three IV assumptions in three languages. Consider the DAG: $\{Z \to T,\ T \to Y,\ U \to T,\ U \to Y,\ X \to T,\ X \to Y,\ X \to Z\}$ with U unobserved.

(a) List all back-door paths from T to Y. Does X alone satisfy the back-door criterion? Explain.
(b) Verify the three IV assumptions using d-separation: (i) Relevance: show Z and T are not d-separated in $\mathcal{G}$. (ii) Exogeneity: show $Z \perp\!\!\!\perp U \mid X$ in $\mathcal{G}$. (iii) Exclusion: show $Y \perp\!\!\!\perp Z \mid T, X$ in $\mathcal{G}_{\overline{T}}$.
(c) Now add the arrow $Z \to Y$ to the DAG. Which IV assumption is violated? Show explicitly which step of the Wald derivation in Section 8.7 breaks down.
(d) Translate each of the three IV assumptions into the structural language: write the equations for T and Y and identify which coefficient restriction corresponds to each assumption.

2. Bias under assumption violations. Let $Y = \beta T + \varepsilon$ and $T = \pi Z + \eta$ with $\mathbb{E}[\varepsilon \mid Z] = 0$ and $\pi \neq 0$.

(a) Starting from $\mathbb{E}[Y \mid Z{=}1] - \mathbb{E}[Y \mid Z{=}0]$, substitute the structural equation for Y and simplify. What role does exogeneity play?
(b) Show that $\mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0] = \pi$ in the linear first-stage model.
(c) Derive the Wald estimand and confirm it equals β.
(d) Now suppose the exclusion restriction fails and $Y = \beta T + \delta Z + \varepsilon$ with $\delta \neq 0$. Derive the probability limit of the Wald estimator and confirm the bias formula from Section 7.4.
(e) Suppose instead that exogeneity fails: $\mathbb{E}[\varepsilon \mid Z] = cZ$ for some constant $c \neq 0$. Derive the probability limit of the Wald estimator and express the bias in terms of c and π. Compare the structure of this bias with the exclusion violation bias.

3. Order, rank, and the limits of overidentification. Consider a model with one endogenous variable T and two instruments Z_1 and Z_2, both satisfying exogeneity and exclusion.

(a) State the order condition and verify it is satisfied.
(b) State the rank condition. What would it mean geometrically if the rank condition failed — i.e., if Z_1 and Z_2 were perfectly collinear in the first-stage regression?
(c) Explain intuitively why having two valid instruments rather than one should improve estimation precision.
(d) Now suppose Z_1 is valid but Z_2 violates the exclusion restriction. Under what conditions does the Sargan–Hansen J-test have power to detect Z_2's invalidity? Under what conditions does the test fail?
(e) Why does passing the J-test not confirm that both Z_1 and Z_2 are valid? Give a concrete example in which both instruments are invalid and the J-test has no power.

4. Compliance types and the LATE. In a binary instrument, binary treatment study, suppose the population has: 30% compliers with average treatment effect $\tau_c = 6$; 25% always-takers with $\tau_a = 3$; 45%

never-takers with $\tau_n = 1$; no defiers.

(a) Compute $P(\text{complier}) = \mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0]$.
(b) Compute the ATE as a weighted average of τ_c, τ_a, τ_n with appropriate weights.
(c) The Wald estimand equals $\tau_c = 6$. By how much does this overstate the ATE, and why?
(d) A second study uses a different binary instrument Z' with a complier population of 50% and a LATE of 2. Is this contradictory? What can you infer about the relative treatment effect in the two complier populations?
(e) Explain, using compliance type language, why the denominator of the Wald estimand equals $P(\text{complier})$.

5. The exclusion restriction: plausibility and violations. Evaluate the exclusion restriction for each proposed instrument. For each, state (i) whether the restriction is plausible and why; (ii) a specific mechanism by which it could be violated; and (iii) whether the violation would bias the IV estimate upward or downward.

(a) **Instrument**: rainfall in the home region of a politician, used as an instrument for government infrastructure spending. **Outcome**: local economic growth.
(b) **Instrument**: distance to the nearest hospital, used as an instrument for hospital admission. **Outcome**: 30-day mortality.
(c) **Instrument**: a randomly assigned financial incentive to enroll in a health screening program. **Outcome**: health status two years later.
(d) **Instrument**: lottery number in the Vietnam-era draft lottery, used as an instrument for military service. **Outcome**: lifetime earnings. [*This is the Angrist (1990) study; discuss why this instrument is widely regarded as satisfying the exclusion restriction.*]

6. IV versus back-door adjustment. A researcher studies the effect of job training (T) on earnings (Y). Two strategies are available: (A) a rich set of pre-treatment covariates X and a propensity-score estimator; (B) a lottery that randomly selected units to be *offered* training (not required to attend), used as instrument Z.

(a) Under what assumption does strategy (A) identify the ATE? What specific unobserved variable would most plausibly violate this assumption?
(b) Strategy (B) identifies a LATE. Describe the complier population in words. Is the LATE likely to be larger or smaller than the ATE in this setting? Explain.
(c) Both strategies are implemented and yield estimates of \$1,800 and \$2,400 per year, respectively. Describe a Hausman-type test that uses both estimates. Under what null hypothesis does the test have an approximate χ^2 distribution?
(d) If the two estimates differ significantly, which strategy would you trust more and why? What additional evidence would help distinguish the two explanations (endogeneity bias in (A) versus LATE $\neq$ ATE in (B))?

Chapter 8

Mediation and Front-Door Identification

Learning Objectives

By the end of this chapter, students should be able to:

1. Explain the distinction between the total causal effect of T on Y and a pathway-specific effect that operates through a mediator M, and describe why this distinction matters scientifically.
2. Draw the prototype mediation DAG, write its structural equations, and identify the direct and indirect pathways.
3. Define the controlled direct effect (CDE) using the do-operator, identify it via the back-door formula for the joint intervention (T, M), and explain why it depends on the fixed level m.
4. Define the natural direct and indirect effects (NDE, NIE) using the potential outcomes notation $Y(t, M(t'))$, state the NDE+NIE=TE decomposition, and explain why these quantities involve cross-world counterfactuals.
5. State the four sequential ignorability assumptions for identification of natural effects, write the mediation formula, and identify which assumption is violated when there is an unmeasured treatment-induced mediator–outcome confounder.
6. Set up the Baron–Kenny three-equation system, derive the product and difference formulas for the indirect effect, and explain why the decomposition fails in nonlinear or interaction models.
7. State the three front-door conditions, derive the front-door formula using the do-calculus, and explain why the front-door graph enables identification despite unobserved T–Y confounding.
8. Contrast mediation analysis and instrumental variables on the dimensions of variable position, identification goal, and key assumption.

8.1 Motivation: Mechanisms

The identification results of Chapters 5–7 all answer the same question: *what is the total causal effect of T on Y?* Mediation analysis asks a finer question: *through what mechanism does that effect operate?*

The total effect of T on Y may flow along multiple causal pathways. Some of this effect passes through an intermediate variable M — the *mediator* — along the path $T \to M \to Y$. The remainder flows directly along $T \to Y$, bypassing the mediator entirely. Mediation analysis aims to study mechanisms by defining direct and indirect effect concepts that target each pathway; only some of these concepts yield an additive decomposition of the total effect.

This mechanism question matters for scientific and policy reasons. In a clinical trial of a behavioral intervention (T) on depression (Y), a researcher may want to know how much of the benefit operates through improved sleep quality (M) versus other pathways — because if sleep is the main channel, targeting sleep directly may be a more efficient intervention. In an economics study of education (T) on wages (Y), how much operates through occupation (M) versus cognitive skills? The answer determines whether a policy should target educational attainment or occupational access.

Running example. Throughout this chapter we anchor the abstract formulas to a single concrete scenario: T is a randomized behavioral intervention, M is self-reported sleep quality measured mid-trial, and Y is a depression score (e.g., on the PHQ-9 scale).

The challenge is that mediators are *post-treatment variables*: they are affected by the treatment, and may themselves be confounded with the outcome. Conditioning on a post-treatment variable creates exactly the collider and selection-bias problems studied in Chapters 2 and 3. A naïve approach — simply including M as a covariate in a regression of Y on T — conflates adjustment with mediation and can introduce bias even in a randomized experiment.

This chapter also develops the *front-door criterion*, a distinct identification strategy that uses the mediation structure of the DAG to identify causal effects even when treatment and outcome are confounded by an unobserved variable, making mediation analysis relevant not only to mechanism research but also to the core identification problem of earlier chapters.

8.2 The Mediation DAG

8.2.1 The Prototype Graph

```
\usetikzlibrary{arrows.meta,positioning}
\definecolor{accent}{RGB}{46,117,182}
\definecolor{defbg}{RGB}{238,244,251}
\definecolor{darkgrey}{RGB}{80,80,80}
\definecolor{causalgreen}{RGB}{26,122,58}
\tikzset{
  node/.style={circle,draw=accent,fill=defbg,thick,minimum size=8mm,font=\small},
  unode/.style={circle,draw=darkgrey,fill=gray!8,dashed,thick,minimum size=8mm,font=\small},
  edge/.style={-{Stealth[length=4pt]},thick,color=accent},
  dedge/.style={-{Stealth[length=4pt]},thick,color=darkgrey,dashed},
  gedge/.style={-{Stealth[length=4pt]},thick,color=causalgreen}
}
\begin{tikzpicture}[node distance=1.9cm]
  \node[node] (T) at (0,0) {$T$};
  \node[node] (M) at (2.4,0) {$M$};
  \node[node] (Y) at (4.8,0) {$Y$};
  \node[node] (X) at (2.4,1.5) {$\mathbf{X}$};
  \node[unode] (U) at (2.4,-1.5) {$U$};
  \draw[edge] (T)--(M); \draw[edge] (M)--(Y); \draw[edge] (T) to[bend right=20] (Y);
  \draw[gedge] (X)--(T); \draw[gedge] (X)--(M); \draw[gedge] (X)--(Y);
  \draw[dedge] (U)--(T); \draw[dedge] (U)--(Y);
\end{tikzpicture}
```

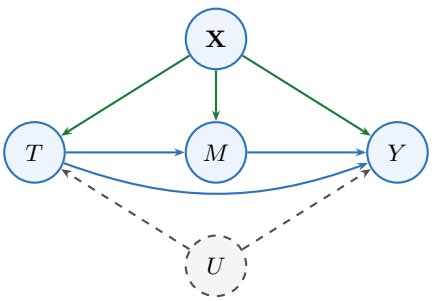


Figure 8.1: The prototype mediation DAG.

The graph encodes two causal pathways: the **direct pathway** $T \to Y$ (treatment affects outcome without passing through the mediator) and the **indirect pathway** $T \to M \to Y$ (treatment first shifts the mediator, which in turn shifts the outcome).

8.2.2 Structural Equations

The prototype graph corresponds to the nonparametric SEM:

$$T = f_T(\mathbf{X}, U, \varepsilon_T), \qquad M = f_M(T, \mathbf{X}, \varepsilon_M), \qquad Y = f_Y(T, M, \mathbf{X}, U, \varepsilon_Y), \tag{8.1}$$

where each arrow corresponds to the presence of the parent in the child's structural equation. U enters both T's and Y's equations, making the confounding paths explicit; U is *absent* from M's equation, reflecting the absence of an arrow $U \to M$ in the DAG.

8.2.3 What Makes Mediation Harder Than Total Effect Estimation

Collider bias. Conditioning on M can open collider paths. Suppose $U \to T$ and $V \to M$ and $V \to Y$, with V unobserved. The path $T \to M \leftarrow V \to Y$ is blocked when M is not conditioned on, but opens as soon as M is included as a covariate. This is precisely why naively regressing Y on (T, M) does not isolate the direct effect.

Mediator–outcome confounding. Even when treatment is randomized, the mediator M is never randomized. An unobserved variable V with $V \to M$ and $V \to Y$ creates a back-door path from M to Y that randomization of T does not close.

Post-Treatment Variables Are Dangerous to Condition On

The mediator M is caused by the treatment T. Conditioning on a post-treatment variable in a regression or matching procedure can open collider paths, introduce selection bias, and produce estimates of neither the total effect nor any well-defined direct effect. This chapter studies three regimes in which conditioning on or marginalizing over M is justified: (1) **CDE** (Section 8.4): both T and M are intervened on via $\text{do}(T, M)$; (2) **NDE/NIE** (Section 8.5–Section 8.6): M is marginalized over using a distribution from a different treatment arm, justified by sequential ignorability; (3) **Front-door** (Section 8.8): M is summed over in the front-door formula, justified by graphical conditions. Outside these three regimes, conditioning on a post-treatment variable should be treated as an error until proven otherwise.

8.2.4 A Working Graph for the Identification Sections

Working Assumption for Sections Section 8.3–Section 8.7

Throughout Section 8.3–Section 8.7, we work with the **reduced prototype graph** obtained from the prototype by absorbing all T–Y and T–M confounding into the observed $\mathbf{X}$ (i.e., U is assumed absent or observed). The front-door Section 8.8 restores U as unobserved, deriving identification of the total effect without access to U.

8.3 Total Causal Effect

Definition: Total Effect

The **total effect** (TE) of T on Y is:

$$\text{TE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y \mid \text{do}(T{=}1)] - \mathbb{E}[Y \mid \text{do}(T{=}0)].$$

Under the working assumption, TE is identified by the back-door formula:

$$\text{TE} = \sum_{\mathbf{x}} [\mathbb{E}[Y \mid T{=}1, \mathbf{X}{=}\mathbf{x}] - \mathbb{E}[Y \mid T{=}0, \mathbf{X}{=}\mathbf{x}]]\, P(\mathbf{X}{=}\mathbf{x}).$$

Running example. The TE is the expected change in depression score if the entire study population were assigned to the intervention versus control. It combines the effect operating through sleep improvement with every other pathway. Mediation analysis asks: of this total, how much is due to sleep?

8.4 Controlled Direct Effect

Definition: Controlled Direct Effect (CDE)

The **controlled direct effect** of changing T from 0 to 1 while fixing $M = m$ by intervention is:

$$\text{CDE}(m) = \mathbb{E}[Y \mid \text{do}(T{=}1), \text{do}(M{=}m)] - \mathbb{E}[Y \mid \text{do}(T{=}0), \text{do}(M{=}m)]. \tag{8.2}$$

Equation 8.2 involves two simultaneous interventions, corresponding to the mutilated graph $\mathcal{G}_{\overline{TM}}$ (all edges into both T and M deleted). Fixing $M = m$ for everyone shuts down the indirect pathway $T \to M \to Y$: any remaining T-to-Y effect flows only through the direct edge.

Running example. CDE(m) at \$m = \$ "poor sleep" is the expected change in depression score comparing intervention to control *if every participant's sleep quality were externally held at the poor-sleep level.* Whether sleep can be externally fixed in a clinical trial is a separate question — which is why the policy interpretation of the CDE in this scenario is strained.

Remark: CDE Depends on m

In general, the direct effect may vary across values of m — this is *effect modification by the mediator.* In a linear additive model, CDE(m) $= \tau'$ for all m, a special property of linearity. When T and M interact, CDEs at different values of m differ.

8.4.1 Identification of the CDE

Theorem: Identification of the CDE

Suppose $\mathbf{Z}$ satisfies the back-door criterion for the joint intervention do(T, M) on Y: (i) $\mathbf{Z}$ contains no descendant of T or M, (ii) $\mathbf{Z}$ blocks every back-door path from $\{T, M\}$ to Y, and (iii) **joint positivity** holds: $P(T{=}t, M{=}m \mid \mathbf{Z}{=}\mathbf{z}) > 0$ for P-almost every $\mathbf{z}$. Then:

$$\mathbb{E}[Y \mid \text{do}(T{=}t), \text{do}(M{=}m)] = \sum_{\mathbf{z}} \mathbb{E}[Y \mid t, m, \mathbf{z}]\, P(\mathbf{z}). \tag{8.3}$$

Under the working assumption, $\mathbf{Z} = \mathbf{X}$ is a valid adjustment set.

Graph surgery and back-door adjustment are distinct steps. Graph surgery *defines* the interventional target by deleting arrows into T and M. Expressing that target as a functional of the observed distribution is a separate step requiring a valid adjustment set $\mathbf{Z}$ in the *original* graph. If U is unobserved and $\mathbf{X}$ alone cannot block the back-door path $T \leftarrow U \to Y$, then Equation 8.3 with $\mathbf{Z} = \mathbf{X}$ does not identify the CDE. Alternative strategies (front-door, IV) are required.

8.4.2 Physical Manipulability and the CDE Does Not Decompose

The CDE is only scientifically meaningful when an intervention to fix M at a specified level m is *physically realizable.* In many substantive settings, the mediator cannot be independently manipulated (education-occupation, behavioral intervention-sleep), making the CDE's policy interpretation strained.

The CDE and some "controlled indirect effect" do *not* sum to the total effect in general. The residual TE $-$ CDE(m) depends on m and has no clean do-calculus expression corresponding to "the indirect pathway." The correct estimands for a pathway decomposition are the NDE and NIE, introduced next.

Natural Effects Are a Strictly Harder Estimand Than the CDE

The CDE is a **single-world estimand**: $\mathbb{E}[Y \mid \text{do}(T{=}t), \text{do}(M{=}m)]$ involves one joint intervention, and identification reduces to the back-door criterion for the pair (T, M).
The NDE and NIE are **cross-world estimands**: $Y(t, M(t'))$ requires an individual to simultaneously inhabit two worlds — the world where $T = t$ (which determines Y) and the world where $T = t'$ (which determines M). When $t \neq t'$, no single experimental intervention can realize both at once.

The do-operator alone cannot express this quantity.
The identification price is steep. The CDE requires only the back-door criterion. The NDE and NIE require four sequential ignorability assumptions, including Assumption 4 (no treatment-induced mediator–outcome confounder), which *cannot* be satisfied by randomizing T and cannot be verified from data.

8.5 Natural Direct and Indirect Effects

8.5.1 Cross-World Counterfactuals

The CDE fixes the mediator by external intervention. A more scientifically natural question is: what is the effect of T on Y that bypasses M when M is held at the value it *would naturally take* under the reference treatment $T = 0$? This requires the *nested potential outcomes* notation $Y(t, M(t'))$, which denotes the outcome observed if T were set to t and M were simultaneously set to the value it would naturally take if T were t'. These are called *cross-world counterfactuals* because t and t' may differ.

Definition: Natural Direct and Indirect Effects [@pearl2001direct]

For binary $T \in \{0, 1\}$:

$$\text{NDE} = \mathbb{E}[Y(1, M(0)) - Y(0, M(0))], \tag{8.4}$$

$$\text{NIE} = \mathbb{E}[Y(1, M(1)) - Y(1, M(0))]. \tag{8.5}$$

The **natural direct effect** (NDE) is the expected change in the outcome when T shifts from 0 to 1, holding the mediator at the value it would naturally take under $T = 0$. The **natural indirect effect** (NIE) is the expected change in the outcome due to the shift in the mediator from $M(0)$ to $M(1)$, holding $T = 1$ fixed.

Running example. The NDE is the part of the depression reduction that comes from cognitive, behavioral, or therapeutic-alliance pathways, not from sleep. The NIE is the complementary piece: how much of the depression reduction is due to the sleep improvement that the intervention itself causes.

The TE = NDE + NIE decomposition:

$$\text{TE} = \text{NDE} + \text{NIE}. \tag{8.6}$$

Proof. NDE + NIE $= \mathbb{E}[Y(1, M(0)) - Y(0, M(0))] + \mathbb{E}[Y(1, M(1)) - Y(1, M(0))] = \mathbb{E}[Y(1, M(1)) - Y(0, M(0))] = \mathbb{E}[Y(1) - Y(0)] = \text{TE}$. □

Remark: Alternative Decomposition

The "starred" decomposition uses $\text{NDE}^* = \mathbb{E}[Y(1, M(1)) - Y(0, M(1))]$ and $\text{NIE}^* = \mathbb{E}[Y(0, M(1)) - Y(0, M(0))]$. Both sum to TE, and the two agree when there is no $T \times M$ interaction. When interaction is present, the gap $\text{NDE} - \text{NDE}^* = \text{NIE}^* - \text{NIE}$ quantifies the discrepancy. This textbook uses the Equation 8.4/Equation 8.5 pair (Pearl-style decomposition) throughout.

Remark: CDE and NDE Coincide under Linearity

In the linear additive Baron–Kenny SEM with no $T \times M$ interaction, the CDE and NDE coincide: NDE $= \tau'$ and NIE $= ab$. When there is a $T \times M$ interaction, NDE $\neq$ CDE and the two concepts address different scientific questions.

8.6 Identification of Natural Effects

8.6.1 Sequential Ignorability

Sequential Ignorability [@imai2010identification]

1. **No unmeasured treatment–outcome confounding (joint form).** $\{Y(t', m), M(t)\} \perp\!\!\!\perp T \mid \mathbf{X}$ for all t, t', m. Graphically: $\mathbf{X}$ blocks all back-door paths from T to Y and from T to M.
2. **No unmeasured treatment–mediator confounding.** $M(t) \perp\!\!\!\perp T \mid \mathbf{X}$ for all t. (Follows from the joint Assumption 1; listed separately to highlight the $T \to M$ sub-problem.)
3. **No unmeasured mediator–outcome confounding given T.** $Y(t', m) \perp\!\!\!\perp M \mid T, \mathbf{X}$ for all t', m. Graphically: $(T, \mathbf{X})$ blocks all back-door paths from M to Y.
4. **No treatment-induced mediator–outcome confounder.** There is no post-treatment variable L such that $T \to L$, $L \to M$, and $L \to Y$.

Positivity Conditions

(P1) Treatment overlap. $P(T{=}t \mid \mathbf{X}{=}\mathbf{x}) > 0$ for all $t \in \{0, 1\}$ and for P-almost every $\mathbf{x}$. Randomization of T secures (P1) by design.
(P2) Mediator overlap across treatment arms. $P(M{=}m \mid T{=}t, \mathbf{X}{=}\mathbf{x}) > 0$ whenever $P(M{=}m \mid T{=}t', \mathbf{X}{=}\mathbf{x}) > 0$, for pairs (t, t') in the mediation formula and P-almost every $\mathbf{x}$. This is a data-dependent requirement that no aspect of experimental assignment guarantees.

Assumptions 1–3 are the natural extensions of the Baron–Kenny conditions to the potential outcomes setting. Assumption 4 is the critical new requirement: it rules out a variable L that is caused by the treatment *and* confounds the mediator–outcome relationship.

What randomization of T does and does not provide. Randomizing T satisfies Assumptions 1 and 2 by design. It does *not* satisfy Assumptions 3 or 4. The mediator M is a post-treatment variable that is never randomized; any unobserved variable V with $V \to M$ and $V \to Y$ violates Assumption 3 regardless of how T was assigned. Assumption 4 is even more demanding: it can be violated by a variable L that is itself *caused* by the treatment.

Randomization of T Does Not Secure Assumption 4

A treatment-induced mediator–outcome confounder L is itself caused by the treatment: the path $T \to L$ is set in motion by the intervention, so no aspect of how T is assigned closes the door on Assumption 4. The absence of such an L must be argued from design, timing, measurement, or subject-matter knowledge. Identification of natural effects is strictly harder than identification of the total effect or the CDE.

8.6.2 The Mediation Formula

Theorem: Mediation Formula [@pearl2001direct]

Under sequential ignorability, positivity (P1)–(P2), and with discrete M and $\mathbf{X}$:

$$\mathbb{E}[Y(t, M(t'))] = \sum_{m} \sum_{\mathbf{x}} \mathbb{E}[Y \mid T{=}t, M{=}m, \mathbf{X}{=}\mathbf{x}]\, P(M{=}m \mid T{=}t', \mathbf{X}{=}\mathbf{x})\, P(\mathbf{X}{=}\mathbf{x}). \tag{8.7}$$

Proof Sketch

The derivation proceeds in five labeled steps.
Step 1. By the law of total expectation and the composition axiom $Y(t, M(t')) = Y(t, m)$ on the

event $\{M(t') = m\}$:

$$\mathbb{E}[Y(t, M(t'))] = \sum_{m,\mathbf{x}} \mathbb{E}[Y(t,m) \mid M(t')=m, \mathbf{X}=\mathbf{x}]\, P(M(t')=m \mid \mathbf{X}=\mathbf{x})\, P(\mathbf{X}=\mathbf{x}).$$

Step 2 (cross-world independence under NPSEM-IE). Assumptions 1–4 jointly imply $Y(t,m) \perp\!\!\!\perp M(t') \mid \mathbf{X}$. The role of Assumption 4 is to ensure $M(t')$ shares no source of variation with $Y(t,m)$ beyond $\mathbf{X}$: if such an L existed, $M(t')$ would inherit L-dependence and $Y(t,m)$ would as well. With cross-world independence: $\mathbb{E}[Y(t,m) \mid M(t')=m, \mathbf{X}=\mathbf{x}] = \mathbb{E}[Y(t,m) \mid \mathbf{X}=\mathbf{x}]$.
Step 3 (Assumptions 1 and 3). By $Y(t,m) \perp\!\!\!\perp T \mid \mathbf{X}$ and $Y(t,m) \perp\!\!\!\perp M \mid T, \mathbf{X}$: $\mathbb{E}[Y(t,m) \mid \mathbf{X}=\mathbf{x}] = \mathbb{E}[Y(t,m) \mid T=t, M=m, \mathbf{X}=\mathbf{x}]$.
Step 4 (consistency for Y). On the event $\{T = t, M = m\}$, $Y(t,m) = Y$: $\mathbb{E}[Y(t,m) \mid T=t, M=m, \mathbf{X}=\mathbf{x}] = \mathbb{E}[Y \mid T=t, M=m, \mathbf{X}=\mathbf{x}]$.
Step 5 (Assumption 2 and consistency for M). $P(M(t')=m \mid \mathbf{X}=\mathbf{x}) = P(M=m \mid T=t', \mathbf{X}=\mathbf{x})$.
Substituting Steps 2–5 into Step 1 yields Equation 8.7. □

Interpretation. The mediation formula "mixes" the outcome regression under $T = t$ with the mediator distribution under $T = t'$. To compute the NDE, set $t = 1$ and $t' = 0$: take the conditional mean of Y at treatment 1, but weight the mediator by its distribution under treatment 0. This counterfactual reweighting is what makes the formula non-trivial. The presence of the second treatment index t' on the right-hand side is the visible trace of the cross-world step.

The NDE and NIE from the formula:

$$\text{NDE} = \sum_{m,\mathbf{x}} [\mathbb{E}[Y \mid 1, m, \mathbf{x}] - \mathbb{E}[Y \mid 0, m, \mathbf{x}]]\, P(M=m \mid T=0, \mathbf{x})\, P(\mathbf{x}),$$

$$\text{NIE} = \sum_{m,\mathbf{x}} \mathbb{E}[Y \mid 1, m, \mathbf{x}]\, [P(M=m \mid T=1, \mathbf{x}) - P(M=m \mid T=0, \mathbf{x})]\, P(\mathbf{x}).$$

Estimation Recipe: Plug-In for the Mediation Formula

Given data $\{(Y_i, T_i, M_i, \mathbf{X}_i)\}_{i=1}^n$, a plug-in estimator requires two working models:

1. **Fit an outcome model.** Regress Y on $(T, M, \mathbf{X})$ to obtain $\hat{\mu}(t, m, \mathbf{x})$.
2. **Fit a mediator model.** Regress M on $(T, \mathbf{X})$ to obtain $\hat{p}(m \mid t, \mathbf{x})$.
3. **Predict outcomes at the evaluation arm.** For each unit i, compute $\hat{\mu}(t, m, \mathbf{X}_i)$ at the evaluation treatment level t.
4. **Reweight or simulate the mediator at the reference arm.** For discrete M, weight $\hat{\mu}(t, m, \mathbf{X}_i)$ by $\hat{p}(m \mid t', \mathbf{X}_i)$.
5. **Average over the empirical distribution of X.** The plug-in estimate is $n^{-1} \sum_i \sum_m \hat{\mu}(t, m, \mathbf{X}_i)\, \hat{p}(m \mid t', \mathbf{X}_i)$.
6. **Quantify uncertainty.** Nonparametric bootstrap gives valid confidence intervals under correct model specification. Influence-function-based estimators (Chapters 10–11) deliver asymptotic normality under doubly-robust conditions.

The plug-in estimator is consistent only when both $\hat{\mu}$ and $\hat{p}$ are correctly specified. The semiparametric estimators developed later in the book are designed to partly shield inference from this sensitivity.

Remark: Three Things to Keep in View

1. **Randomization handles two of the four assumptions, not all four.** Randomizing T secures Assumptions 1 and 2 by design. It leaves Assumptions 3 and 4 entirely open.
2. **The formula looks like routine adjustment but is not.** The expression $\sum_m \mathbb{E}[Y \mid t, m, \mathbf{x}]\, P(M=m \mid t', \mathbf{x})$ resembles a standardization formula, but it computes a cross-world expectation $\mathbb{E}[Y(t, M(t'))]$, not a do-expression.
3. **Sequential ignorability is an untestable assumption bundle.** Unlike treatment ignorability, the mediator–outcome ignorability in Assumption 3 and the no-treatment-induced-confounder condition in Assumption 4 must be defended on subject-matter grounds in every application.

8.7 The Linear Mediation Model: A Historical Special Case

8.7.1 The Baron–Kenny Three-Equation System

The regression-based approach of Baron and Kenny (1986) restricts the reduced prototype graph to a linear SEM:

$$Y = \alpha_1 + \tau T + \gamma_1^\top \mathbf{X} + \varepsilon_1, \tag{8.8}$$

$$M = \alpha_2 + aT + \gamma_2^\top \mathbf{X} + \varepsilon_2, \tag{8.9}$$

$$Y = \alpha_3 + \tau' T + bM + \gamma_3^\top \mathbf{X} + \varepsilon_3. \tag{8.10}$$

The four coefficients: τ (total effect), a (first-stage effect of T on M), τ' (direct effect of T on Y controlling for M), b (second-stage effect of M on Y controlling for T).

8.7.2 The Component Pathways

First stage ($T \to M$): Equation Equation 8.9 implements the back-door formula for T on M. Under Condition 2, conditioning on $\mathbf{X}$ blocks all back-door paths from T to M, and a identifies $\mathbb{E}[M(1)] - \mathbb{E}[M(0)]$.

Second stage ($M \to Y$ given T): Equation Equation 8.10 implements the back-door formula for M on Y given T. Conditioning on $(T, \mathbf{X})$ blocks every back-door path from M to Y in the reduced prototype graph, *provided Assumption 3 (no unmeasured M–Y confounding given T) holds.*

> The Second Stage Is Harder Than It Looks
>
> Even in a randomized experiment, the mediator M is never randomized. An unobserved variable V with $V \to M$ and $V \to Y$ creates a back-door path from M to Y that conditioning on $(T, \mathbf{X})$ cannot block. Identifying b requires no-unmeasured-confounding for the mediator–outcome relationship, an assumption that randomization of T does not provide.

8.7.3 The Product and Difference Formulas

> Proposition: Mediation Decomposition in the Linear Model [@baron1986moderator]
>
> Under equations Equation 8.8–Equation 8.10: $\tau = \tau' + ab$.
> The indirect and direct effects are:
>
> $$\tau_{\text{ind}} = ab \quad \text{(product method)}, \qquad \tau_{\text{dir}} = \tau - ab = \tau' \quad \text{(difference method)}. \tag{8.11}$$

Proof. Substitute Equation 8.9 into Equation 8.10: $Y = (\alpha_3 + b\alpha_2) + (\tau' + ab)T + (\gamma_3 + b\gamma_2)^\top \mathbf{X} + (b\varepsilon_2 + \varepsilon_3)$. Comparing with Equation 8.8 gives $\tau = \tau' + ab$. □

> The Decomposition $\tau = \tau' + ab$ Is an Algebraic Identity, Not a Causal Theorem
>
> The equality holds because linearity makes the indirect effect separable and additive. It does **not** hold in general:
> - In **nonlinear models** (binary outcomes, count outcomes, survival models), the product and difference methods yield numerically different estimates. Neither equals the NIE in general.
> - When T **and** M **interact** in their effect on Y, the CDE depends on m, the NDE and CDE diverge, and the indirect effect cannot be summarized by a single number ab.
> - For **non-continuous mediators**, the product ab has no simple causal interpretation outside the linear normal model.
>
> The correct generalization is the mediation formula (Equation 8.7), which reduces to ab and τ' only in the linear, no-interaction special case.

Running example. Suppose a randomized trial yields $\hat{\tau} = 0.50$, $\hat{a} = 0.40$, $\hat{b} = 0.60$, $\hat{\tau}' = 0.26$.

- Indirect effect (product method): $\hat{a}\hat{b} = 0.40 \times 0.60 = 0.24$.
- Direct effect (difference method): $\hat{\tau} - \hat{a}\hat{b} = 0.50 - 0.24 = 0.26 = \hat{\tau}'$.

- Proportion mediated: $\hat{a}\hat{b}/\hat{\tau} = 0.24/0.50 = 0.48$ — roughly 48% of the total effect operates through sleep improvement.

Effect	Formula	Path(s)
Total	τ	$T \to Y$ and $T \to M \to Y$ combined
Direct	$\tau' = \tau - ab$	$T \to Y$ only
Indirect	ab	$T \to M \to Y$ only
Proportion mediated	ab/τ	Share of total effect via M

8.7.4 Inference: The Sobel Test and Bootstrap

The delta method gives an approximate variance for the product $\hat{a}\hat{b}$:

$$\widehat{\mathrm{Var}}(\hat{a}\hat{b}) \approx \hat{b}^2\,\widehat{\mathrm{Var}}(\hat{a}) + \hat{a}^2\,\widehat{\mathrm{Var}}(\hat{b}) + 2\,\hat{a}\,\hat{b}\,\widehat{\mathrm{Cov}}(\hat{a},\hat{b}). \tag{8.12}$$

The **Sobel test** (Sobel 1982) drops the cross-covariance:

$$\widehat{\mathrm{Var}}_{\mathrm{Sobel}}(\hat{a}\hat{b}) = \hat{b}^2\,\widehat{\mathrm{Var}}(\hat{a}) + \hat{a}^2\,\widehat{\mathrm{Var}}(\hat{b}), \tag{8.13}$$

yielding $z = \hat{a}\hat{b}/\sqrt{\widehat{\mathrm{Var}}_{\mathrm{Sobel}}(\hat{a}\hat{b})}$. In practice, bootstrap confidence intervals for ab are preferred over the Sobel test because the distribution of a product of estimates is skewed in finite samples.

> The "Significance of Both Paths" Criterion Is Not a Test for Mediation
>
> A common misuse declares mediation when (i) $T \to Y$ is significant, (ii) $T \to M$ is significant, (iii) $M \to Y$ is significant. This approach has three defects: (1) statistical significance $\neq$ mediation; (2) the Sobel test tests a regression product, not a causal quantity; (3) zero total effect does not preclude indirect effects (direct and indirect effects of opposite sign can cancel). The modern alternative is to estimate ab or the NIE directly, construct bootstrap confidence intervals, and interpret as a point estimate with uncertainty.

8.7.5 The Baron–Kenny Assumptions: Two Distinct Categories

Causal ignorability conditions (conditions 1–3): these are causal identification assumptions about unmeasured confounding. Violating them introduces bias that no amount of additional data can remove.

1. No unmeasured T–Y confounding: $\varepsilon_1 \perp\!\!\!\perp T \mid \mathbf{X}$.
2. No unmeasured T–M confounding: $\varepsilon_2 \perp\!\!\!\perp T \mid \mathbf{X}$.
3. No unmeasured M–Y confounding given T: $\varepsilon_3 \perp\!\!\!\perp M \mid T, \mathbf{X}$.

Structural modeling restrictions (conditions 4–5): these are parametric assumptions about functional form. Violating them does not introduce identification bias in the causal sense, but ab and τ' no longer equal the NDE and NIE.

4. Linearity and additivity: equations Equation 8.8–Equation 8.10 are correctly specified as linear and additive.
5. No $T \times M$ interaction: the coefficient b is the same for all values of T.

Conditions 1–3 cannot be tested from observed data at all. Conditions 4–5 can be partially probed by residual diagnostics and interaction terms.

> Framework Map
>
Estimand	Framework needed	Section
Total effect $P(y \mid \mathrm{do}(t))$	Do-calculus	Ch. 5 (back-door)
Controlled direct effect (CDE)	Do-calculus	Section 8.4

Natural direct effect (NDE)	Potential outcomes	Section 8.5
Natural indirect effect (NIE)	Potential outcomes	Section 8.5
Total effect via front-door	Do-calculus	Section 8.8

The first and last rows name the same estimand — the total effect — but identified by different routes.

8.8 Front-Door Identification

8.8.1 The Front-Door DAG

Every identification strategy in Section 8.4–Section 8.6 assumed an observed covariate set $\mathbf{X}$ that blocks the back-door paths from T to Y through U. What if U is wholly unobserved and no such adjustment set exists? The front-door criterion turns this obstacle into an opportunity: under two additional restrictions on the prototype mediation graph, the mediation structure itself provides identification of the *total effect* without conditioning on U.

The two restrictions are: (i) remove the direct $T \to Y$ edge, so M fully mediates; and (ii) require U has no arrow into M, so the $T \to M$ sub-effect is unconfounded.

Why the running example does not apply here. The depression/sleep scenario fails Condition 1 (full mediation): a behavioral intervention plausibly operates through several non-sleep channels. The canonical front-door example is Pearl's smoking–tar–cancer graph: T = smoking, M = tar deposits, Y = lung cancer, U = genetic susceptibility. If all of smoking's carcinogenic effect flows through tar, and genetic susceptibility does not act on tar directly, the front-door formula identifies the causal effect without observing U.

```
\usetikzlibrary{arrows.meta,positioning}
\definecolor{accent}{RGB}{46,117,182}
\definecolor{defbg}{RGB}{238,244,251}
\definecolor{darkgrey}{RGB}{80,80,80}
\tikzset{
  node/.style={circle,draw=accent,fill=defbg,thick,minimum size=8mm,font=\small},
  unode/.style={circle,draw=darkgrey,fill=gray!8,dashed,thick,minimum size=8mm,font=\small},
  edge/.style={-{Stealth[length=4pt]},thick,color=accent},
  dedge/.style={-{Stealth[length=4pt]},thick,color=darkgrey,dashed}
}
\begin{tikzpicture}[node distance=1.9cm]
  \node[node] (T) at (0,0) {$T$};
  \node[node] (M) at (2.8,0) {$M$};
  \node[node] (Y) at (5.6,0) {$Y$};
  \node[unode] (U) at (2.8,1.8) {$U$};
  \draw[edge] (T)--(M); \draw[edge] (M)--(Y);
  \draw[dedge] (U)--(T); \draw[dedge] (U)--(Y);
\end{tikzpicture}
```

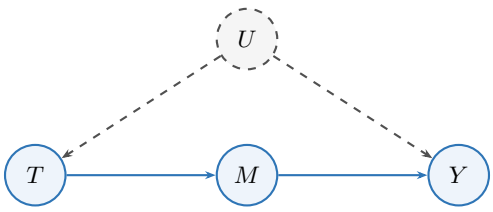


Figure 8.2: The front-door graph: no direct $T \to Y$ edge, and U has no arrow into M.

Two Different Questions About a Mediator-Like Variable

	Ordinary mediation	**Front-door identification**
Question	How much of the total effect of T on Y flows through M?	Can M be used to identify the total effect despite unmeasured T–Y confounding?
Role of M	Pathway: carries part of the causal effect	Relay: routes around unmeasured T–Y confounding
Target estimand	NDE, NIE (decomposition)	$P(y \mid \text{do}(t))$ (total effect)
Direct $T \to Y$ edge	Present	Absent (required)

The front-door formula does not decompose the total effect — it identifies the total effect as a whole, using M as a relay that is unconfounded on the T-side.

"Instrument-Like" Does Not Mean Instrument

The front-door mediator M lies on the causal path from T to Y and enters the Y structural equation directly. The IV instrument Z satisfies the exclusion restriction — it affects Y only through T. These are structurally opposite positions. The two strategies share the goal of identifying T's total effect under unobserved confounding, but they place the third variable in fundamentally different roles.

8.8.2 The Three Front-Door Conditions

Front-Door Conditions

A variable M satisfies the **front-door criterion** for the effect of T on Y if:

1. **Full mediation.** All directed paths from T to Y pass through M (no direct $T \to Y$ edge).
2. **No unblocked back-door path from T to M.** All back-door paths from T to M are blocked (no unobserved variables affect both T and M).
3. **No unblocked back-door path from M to Y given T.** All back-door paths from M to Y are blocked by conditioning on T.

In the front-door DAG: Condition 1 holds (no $T \to Y$ edge). Condition 2 holds (U has no arrow into M). Condition 3 holds: the only back-door path from M to Y is $M \leftarrow T \leftarrow U \to Y$, which is blocked by conditioning on T.

8.8.3 Derivation of the Front-Door Formula

Theorem: Front-Door Formula [@pearl1995causal]

Suppose M satisfies the front-door criterion and positivity conditions (F1) $P(T{=}t') > 0$ for every t' and (F2) for every m with $P(M{=}m \mid T{=}t) > 0$ and t' with $P(T{=}t') > 0$, $P(M{=}m \mid T{=}t') > 0$. Then:

$$P(y \mid \text{do}(T{=}t)) = \sum_{m} P(m \mid t) \sum_{t'} P(y \mid m, t')\, P(t'). \tag{8.14}$$

Proof

Step 1: Identify the effect of T on M. By Condition 2, there are no unblocked back-door paths from T to M. The empty set is a valid back-door adjustment set, so:

$$P(m \mid \text{do}(T{=}t)) = P(m \mid t).$$

Step 2: Identify the effect of M on Y. By Condition 3, conditioning on T blocks all back-door

paths from M to Y. The set $\{T\}$ is a valid back-door adjustment set:

$$P(y \mid \mathrm{do}(M{=}m)) = \sum_{t'} P(y \mid m, t')\, P(t').$$

Step 3: Combine via full mediation. The law of total probability under $\mathrm{do}(T{=}t)$ gives:

$$P(y \mid \mathrm{do}(T{=}t)) = \sum_{m} P(m \mid \mathrm{do}(T{=}t))\, P(y \mid \mathrm{do}(T{=}t), \mathrm{do}(M{=}m)).$$

By Condition 1 (no direct $T \to Y$ edge), once $M = m$ is fixed by intervention, T is d-separated from Y in $\mathcal{G}_{\overline{T}\,\overline{M}}$. By Rule 3 of the do-calculus, $P(y \mid \mathrm{do}(T{=}t), \mathrm{do}(M{=}m)) = P(y \mid \mathrm{do}(M{=}m))$. Substituting Steps 1 and 2 gives Equation 8.14. □

Remark: Two Unconfounded Sub-Effects

The front-door formula achieves identification in two steps: (1) T **to** M: there is no confounding on the $T \to M$ edge (U does not affect M), so $P(m \mid t)$ is the causal effect. (2) M **to** Y: there is back-door confounding on $M \to Y$ through $M \leftarrow T \leftarrow U \to Y$, but T is a non-collider, so conditioning on T closes it. The result $P(y \mid m, t')$ is then averaged over the marginal distribution of T.
The key insight is that U confounds T and Y but *not* the $T \to M$ edge. The front-door formula exploits this asymmetry without ever observing or conditioning on U.

8.9 Mediation vs. Instrumental Variables

Feature	**Instrumental Variables**	**Mediation Analysis**
Position of third variable	Pre-treatment (Z precedes T)	Post-treatment (M follows T)
Causal role	Exogenous source of variation in T	Pathway through which T affects Y
Primary goal	Identification of $T \to Y$ effect	Mechanism analysis (decomposition)
Key assumption	Exclusion: Z affects Y only through T	Sequential ignorability: no unmeasured M–Y confounding
Estimand	LATE (Wald) or ATE (homogeneity)	NDE, NIE, or CDE
Unobserved T–Y confounders	Permitted	Must be addressed separately
Testability	Relevance testable; exclusion untestable	Sequential ignorability untestable

The Key Conceptual Distinction

	IV	**Mediation**
What the third variable does	Generates clean variation in T (exogenous source)	Carries part of T's causal effect (pathway)
Exclusion vs. inclusion	Z excluded from Y's structural equation	M included in Y's structural equation
Question answered	Does T cause Y?	How does T cause Y?

Can the same variable be both? Not for the same treatment–outcome relation. Within a single causal question of how T affects Y, the mediator role places M *on* the causal path (inclusion required), whereas the IV role demands the exclusion restriction. These are mutually incompatible structural assumptions. The front-door identification formula is the closest bridge between the two within a single (T, Y) analysis:

it uses the mediator M to identify the total effect of T on Y even when T is confounded — but the front-door M is not an instrument.

8.10 Chapter Summary

Symbol	Meaning
TE	Total effect $\mathbb{E}[Y(1) - Y(0)]$
CDE(m)	Controlled direct effect at mediator level m Equation 8.2
NDE	Natural direct effect Equation 8.4
NIE	Natural indirect effect Equation 8.5
$Y(t, M(t'))$	Cross-world counterfactual (nested potential outcome)
$\tau = \tau' + ab$	Baron–Kenny decomposition (linear model only)
Equation 8.7	Mediation formula (nonparametric)
Equation 8.14	Front-door formula

1. **Mediation studies mechanisms.** Mediation analysis defines direct and indirect effect concepts that target the pathways $T \to Y$ and $T \to M \to Y$. Only natural effects yield an additive TE decomposition; the CDE does not.
2. **The total effect is the baseline estimand.** TE $= \mathbb{E}[Y(1) - Y(0)]$ captures all pathways. It may be identified by randomization, back-door adjustment, front-door identification, or IV.
3. **The CDE uses do-calculus.** The CDE fixes $M = m$ by joint intervention $\text{do}(T, M)$ and is identified by the back-door formula applied to the pair (T, M). The CDE depends on the fixed level m and has no natural "indirect" complement.
4. **Natural effects require potential outcomes.** The NDE and NIE involve cross-world counterfactuals $Y(t, M(t'))$ that cannot be expressed with the do-operator alone. They are identified by the mediation formula under sequential ignorability. The critical Assumption 4 (no treatment-induced mediator–outcome confounder) is not secured by randomization of T and must be defended on subject-matter grounds.
5. **The linear model simplifies but restricts.** The Baron–Kenny system gives $\tau_{\text{ind}} = ab$ and $\tau_{\text{dir}} = \tau'$, with $\tau = \tau' + ab$. This decomposition is purely algebraic and holds only under linearity and no interaction.
6. **Front-door identification uses mediation structure.** When M fully mediates $T \to Y$, no $T \to M$ confounding exists, and T blocks the back-door paths from M to Y, the front-door formula identifies the total effect despite unobserved T–Y confounding, by composing two unconfounded sub-effects.
7. **Mediation and IV are complementary, not equivalent.** IV uses a pre-treatment variable to generate exogenous variation in T; mediation uses a post-treatment variable to study how the causal effect operates. The same variable cannot simultaneously serve as a mediator and a valid IV for the same treatment–outcome relation.

8.11 Problems

1. Identifying the CDE. Consider the DAG: $T \to M$, $T \to Y$, $M \to Y$, $X \to T$, $X \to M$, $X \to Y$, with all variables observed.

(a) Write the identification formula for $\mathbb{E}[Y \mid \text{do}(T{=}1), \text{do}(M{=}m)]$ using the back-door criterion for the joint intervention (T, M).
(b) Add an unobserved U with $U \to T$ and $U \to Y$. Does the back-door formula still identify the CDE? Explain which condition fails.
(c) Instead add U with $U \to M$ and $U \to Y$. Does the back-door formula still identify the CDE? Explain.

2. CDE vs. total effect. In the reduced prototype graph, let $\mathbf{Z}$ satisfy the back-door criterion for both the total effect and the joint intervention (T, M).

(a) Write expressions for the total effect and CDE(m) using the back-door formula.
(b) Under what graphical condition does CDE(m) equal the total effect for all m? Interpret this condition.

3. Natural direct and indirect effects. Verify the NDE + NIE = TE decomposition algebraically for the linear SEM $M = \alpha T + \eta$, $Y = \beta T + \gamma M + \varepsilon$ (no interaction).

(a) Compute $Y(t, M(t'))$ in the linear model. Show that $Y(t, M(t')) = \beta t + \gamma(\alpha t') + \text{noise}$.
(b) Derive NDE $= \beta$ and NIE $= \alpha\gamma$ from definitions Equation 8.4–Equation 8.5.
(c) Confirm NDE + NIE $= \beta + \alpha\gamma =$ TE.
(d) Now suppose a $T \times M$ interaction is added: $Y = \beta T + \gamma M + \delta(T \cdot M) + \varepsilon$. Compute CDE($m$) and NDE. Show that NDE $\neq$ CDE(m) when $\delta \neq 0$.

4. The Baron–Kenny three-equation system. In the reduced prototype graph with the linear SEM Equation 8.8–Equation 8.10:

(a) State the three identification assumptions. For each, give the graphical condition in terms of back-door paths.
(b) Derive the equality $\tau = \tau' + ab$ algebraically.
(c) Suppose $\hat{\tau} = 0.50$, $\hat{a} = 0.40$, $\hat{b} = 0.60$, $\hat{\tau}' = 0.26$. Compute the indirect effect by both the product and difference methods. Do they agree? Compute the proportion mediated.
(d) With $\widehat{\text{SE}}(\hat{a}) = 0.08$ and $\widehat{\text{SE}}(\hat{b}) = 0.10$, compute the Sobel standard error for $\hat{a}\hat{b}$ using Equation 8.13 and construct an approximate 95% confidence interval.

5. The critical role of Assumption 3. Consider the graph where an unobserved V has $V \to M$ and $V \to Y$, with T randomized.

(a) Identify all back-door paths from M to Y in this graph.
(b) Can any combination of observed variables $(T, \mathbf{X})$ block all of these paths? Explain using d-separation.
(c) Suppose an analyst fits Equation 8.10 ignoring V and obtains $\hat{b} = 0.80$. In which direction is $\hat{b}$ biased if V has positive effects on both M and Y?
(d) State the additional data structure that would be needed to identify the second-stage effect non-parametrically.

6. Front-door identification. Consider the front-door graph.

(a) Verify that the three front-door conditions hold.
(b) Walk through the three-step proof of the Front-Door Formula (Equation 8.14): identify which do-calculus rule justifies each step.
(c) Add a direct edge $T \to Y$ to the graph. Which front-door condition is violated? Does formula Equation 8.14 still hold?
(d) Explain why the front-door graph does not require no unmeasured M–Y confounding given T as an identifying assumption.

7. Mediation vs. instrumental variables. A researcher studies the effect of a job training program (T) on wages (Y). She proposes two intermediate variables: (A) motivation (M_A), measured after the program starts; (B) a lottery that randomly selects applicants for admission (Z), measured before the program.

(a) For variable (A): draw the mediation DAG including M_A, T, Y, and unobserved ability U. State the sequential ignorability assumption needed to identify the NIE through M_A. Explain why randomization of T does not automatically satisfy this assumption.
(b) For variable (B): draw the IV DAG with Z, T, Y, and U. State the three IV assumptions. Explain why the exclusion restriction and the "mediator inclusion" of mediation analysis are mutually incompatible conditions for the same intermediate variable.
(c) The researcher argues that M_A (motivation) and Z (lottery) are both "intermediate" variables and that the analyses are interchangeable. Write a one-paragraph critique of this argument, using the distinctions from Section 8.9.
(d) Can the front-door formula be applied if motivation M_A fully mediates the effect of T on Y and U (unobserved ability) does not directly affect M_A? State the three conditions and assess whether they hold.

Chapter 9

Sensitivity Analysis and Partial Identification

Learning Objectives

By the end of this chapter, students should be able to:

1. Distinguish sampling uncertainty, model misspecification, and identification uncertainty, and explain why confidence intervals do not measure the credibility of identifying assumptions.
2. Formulate a sensitivity analysis by introducing a sensitivity parameter λ and reporting the sensitivity curve, bounds, or tipping point associated with it.
3. Carry out a sensitivity analysis for unmeasured confounding under back-door adjustment, including the linear bias decomposition and the binary-confounder formula.
4. State and use the three canonical sensitivity models: the Rosenbaum Γ-sensitivity model, the E-value of VanderWeele and Ding (2017), and the marginal sensitivity model of Tan (2006) and Zhao et al. (2019).
5. Interpret sensitivity parameters through benchmarking with observed covariates.
6. State and prove Manski's no-assumption bound on the average treatment effect, and contrast point identification with partial identification.
7. Recognize positivity and overlap violations as a form of identification failure, and interpret trimming as a change of estimand.
8. Explain why weak-instrument diagnostics do not address exogeneity or exclusion violations, and apply the IV bias formula for a scalar instrument.
9. Describe how sensitivity analysis applies to mediation assumptions through a residual-correlation parameter.
10. Recognize sensitivity analysis as complementary to the modern estimation methods of Chapters 10–13: orthogonality and cross-fitting protect against nuisance-estimation error, not against violations of causal identification assumptions.

9.1 Why Sensitivity Analysis?

Chapters 5–8 developed a sequence of identification strategies: back-door adjustment, propensity-score re-weighting, instrumental variables, and front-door and mediation analysis. Each strategy identifies a causal parameter as a functional of the observed-data distribution under a corresponding set of assumptions. Each set of assumptions is, however, stated in terms of unobserved quantities — potential outcomes, hypothetical interventions, or unmeasured variables — and cannot be verified from the data alone.

The chapter ahead (Chapter 10) opens by noting that *identification does not by itself provide a statistically reliable estimator*: once a parameter is identified, a separate theory of estimation and inference is still needed. The present chapter makes the complementary point, which closes out Part II:

> *Statistical reliability does not by itself validate identification.*

A confidence interval reports sampling variation *conditional on* a maintained identification assumption. It is silent about whether that assumption is correct. A causal analysis is credible only if both components — identification and estimation — are credible. Sensitivity analysis is the tool for reporting how much the conclusion depends on the first.

9.1.1 Sampling Uncertainty vs. Identification Uncertainty

Consider an analyst who reports $\hat{\tau} = 2.0$, 95% CI $= [1.2,\ 2.8]$. This interval answers one question and only one:

> *If the identifying assumptions are correct, what range of values of τ is consistent with the observed sampling variation?*

As $n \to \infty$ the interval will shrink around τ, provided the identifying assumptions hold. If the assumptions do not hold, the interval will instead shrink around a biased target. More data drawn from the same observational regime do not remove this bias.

Formally, there are three distinct sources of error:

- **Sampling uncertainty:** $\hat{P}_n \neq P$. The empirical distribution differs from the population. Vanishes as $n \to \infty$.
- **Model misspecification:** the working parametric or semiparametric model for a nuisance function is incorrect. Partly addressable by flexible modeling, doubly robust construction, and cross-fitting (Chapters 10–12).
- **Identification uncertainty:** the causal parameter ψ is not in fact identified by the functional $\Psi(P)$ assumed by the analysis. Does not vanish merely by increasing the sample size from the same observational regime; can be reduced only by adding new assumptions, new design information, or new measurements.

The usual confidence interval addresses only the first; the estimation chapters to come address primarily the second; this chapter addresses the third.

> Statistical Significance and Causal Credibility
>
> A narrow, highly significant confidence interval is not evidence that the identifying assumptions are correct. It is evidence that, *if they are correct*, ψ is close to the estimate. A sensitivity analysis is required to assess the second "if."

9.1.2 The Role of Sensitivity Analysis

Sensitivity analysis is a structured way to vary the strength of violations of identifying assumptions and to report how the causal conclusion responds. It is not a test of the assumptions — they involve unobserved quantities and are therefore not testable. It is a statement of *conditional robustness*: how far can the assumptions be violated before the substantive conclusion changes?

Three reporting objects:

- *Sensitivity curve.* A plot of the estimand against a single sensitivity parameter λ, with $\lambda = 0$ recovering the baseline identifying assumption.
- *Sensitivity bounds.* The range of estimand values consistent with any λ in a specified plausibility set Λ.
- *Tipping point.* The smallest violation magnitude sufficient to change the sign, magnitude, or statistical significance of the conclusion.

9.2 A General Framework for Sensitivity Analysis

The three views of sensitivity analysis — curves, bounds, tipping points — all arise from a single construction. Let ψ be the causal estimand and A_0 the baseline identifying assumption with $\psi = \Psi(P)$. A **sensitivity model** is a one-parameter relaxation $\{A_\lambda : \lambda \in \Lambda\}$ of A_0, with A_0 recovered at $\lambda = 0$. Under A_λ:

$$\psi(\lambda) = \Psi(P;\, \lambda), \qquad \Psi(P;\, 0) = \Psi(P). \tag{9.1}$$

The key reporting object is the set $\{\psi(\lambda) : \lambda \in \Lambda\}$.

Definition: Sensitivity Model

A *sensitivity model* for the causal estimand ψ and baseline identifying assumption A_0 is a pair $(\{A_\lambda\}_{\lambda\in\Lambda}, \Psi(\cdot;\cdot))$ such that (i) A_0 is recovered at $\lambda = 0$; (ii) under A_λ and the observable restrictions, $\psi = \Psi(P;\lambda)$; and (iii) $\Psi(\cdot;\lambda)$ is a continuous function of λ.

Sensitivity parameters used in this chapter:

- λ = strength of unmeasured confounding (odds-ratio bound, E-value, or likelihood-ratio bound).
- δ = magnitude of a structural outcome-effect coefficient treated as zero under the baseline assumption (used both for U-outcome effect and for IV exclusion violations).
- α = propensity-score trimming threshold.
- ρ = residual correlation between mediator and outcome disturbances.

Sensitivity curve: $\lambda \mapsto \hat{\psi}(\lambda)$. Useful when the reader wants to see the effect deform as the assumption is relaxed.

Sensitivity bounds: $\psi_L = \inf_\Lambda \psi(\lambda)$, $\psi_U = \sup_\Lambda \psi(\lambda)$.

Tipping point: $\lambda^\star = \inf\{\lambda \in \Lambda : \psi(\lambda) = 0\}$, or $\lambda^\star_{\text{CI}} = \inf\{\lambda : 0 \in \widehat{\text{CI}}(\lambda)\}$.

Example: Tipping-Point Reading

Suppose $\hat{\psi}(0) = 1.8$ with 95% CI $[1.1, 2.5]$, and the sensitivity model produces $\hat{\psi}(\lambda) = 1.8 - 2.0\lambda$. Then: the sensitivity curve is linear in λ; the tipping point for the sign is $\lambda^\star = 0.9$; and the tipping point for significance is $\lambda^\star_{\text{CI}} = 0.55$ (the value at which the lower CI endpoint $1.1 - 2.0\lambda$ hits zero). The significance tipping point is always at least as strict as the sign tipping point.

9.3 Sensitivity to Unmeasured Confounding

The most common application is to unmeasured confounding under the back-door framework. The baseline assumption is conditional exchangeability $Y(t) \perp\!\!\!\perp T \mid X$ together with consistency and positivity. Sensitivity analysis contemplates a world with an unobserved U such that $Y(t) \perp\!\!\!\perp T \mid X, U$ but $Y(t) \not\perp\!\!\!\perp T \mid X$.

9.3.1 The Bias Decomposition

Let $\Delta_{\text{obs}} = \mathbb{E}\{\mathbb{E}(Y \mid T = 1, X) - \mathbb{E}(Y \mid T = 0, X)\}$ denote the observed adjusted contrast. Under ignorability given X alone, $\Delta_{\text{obs}} = \tau$. Under the relaxed condition, the contrast departs from τ by a bias B.

Theorem: Bias Decomposition under Unmeasured Confounding

Suppose $Y(t) \perp\!\!\!\perp T \mid X, U$ holds along with consistency and positivity. Define $m_t(x,u) = \mathbb{E}(Y \mid T = t, X = x, U = u)$ and $g_t(x,u) = P(u \mid T = t, X = x) - P(u \mid X = x)$. Then:

$$\Delta_{\text{obs}} = \tau + B, \tag{9.2}$$

where the *confounding bias* is:

$$B = \mathbb{E}\left\{\sum_t (-1)^{1-t} \int m_t(X,u)\, g_t(X,u)\, du\right\}. \tag{9.3}$$

Proof

Under conditional exchangeability given (X, U) and consistency:

$$\mathbb{E}[Y(t)] = \mathbb{E}\Big\{\int m_t(X,u)\, P(u \mid X)\, du\Big\}.$$

The conditional mean given (T, X) marginalizes U over its distribution *conditional on* T:

$$\mathbb{E}(Y \mid T=t, X) = \int m_t(X,u)\, P(u \mid T=t, X)\, du.$$

Taking expectations over X and subtracting:

$$\Delta_{\text{obs}} - \tau = \mathbb{E}\Big\{\int m_1(X,u)\,[P(u \mid T=1,X) - P(u \mid X)]\, du\Big\} - \mathbb{E}\Big\{\int m_0(X,u)\,[P(u \mid T=0,X) - P(u \mid X)]\, du\Big\} = B. \quad \square$$

Remark: Bias in Words

The observed adjusted contrast is the true effect plus a term controlled by (*the U-outcome association*) $\times$ (*the U-treatment imbalance*). Every sensitivity model in Section 9.4 is a formal version of this slogan.

9.3.2 A Simple Linear Sensitivity Model

In the fully linear model $Y = \alpha + \tau T + \gamma^\top X + \delta U + \varepsilon$ with $T = h(X, U, \eta)$:

Lemma: Linear Sensitivity Bias

Under the linear outcome model and conditional exchangeability given (X, U):

$$B = \delta \cdot \mathbb{E}\{\mathbb{E}(U \mid T=1, X) - \mathbb{E}(U \mid T=0, X)\}. \tag{9.4}$$

Proof. In the linear model, $m_t(x,u) = \alpha + \tau t + \gamma^\top x + \delta u$. The components of m_t not depending on u multiply integrals of $g_t(X,u)$, which integrate to zero. The δu term gives Equation 9.4. $\square$

This gives the canonical teaching decomposition:

$$\text{bias} \approx \underbrace{\delta}_{U\text{-outcome effect}} \times \underbrace{\mathbb{E}\{\mathbb{E}(U \mid T=1, X) - \mathbb{E}(U \mid T=0, X)\}}_{U\text{-treatment imbalance}}. \tag{9.5}$$

Both factors are sensitivity parameters because U is unobserved.

9.3.3 Binary Unmeasured Confounder

When $U \in \{0,1\}$, define $p_t(x) = P(U=1 \mid T=t, X=x)$.

Lemma: Binary-Confounder Bias

If $U \in \{0,1\}$ and the U-outcome contrast is constant across (t,x) at value δ, then:

$$B = \delta \cdot \mathbb{E}\{p_1(X) - p_0(X)\}. \tag{9.6}$$

Proof. With $U \in \{0,1\}$, $m_t(x,u) = m_t(x,0) + u\,\delta$. Substituting into Equation 9.3: $B = \delta\, \mathbb{E}\{P(U{=}1 \mid T{=}1, X) - P(U{=}1 \mid T{=}0, X)\}$. $\square$

Formula Equation 9.6 is the workhorse of applied sensitivity analysis. A sensitivity analysis consists of computing $\hat{\tau}_{\text{sens}} = \hat{\Delta}_{\text{obs}} - \hat{B}$ over a grid of $(\delta,\ \mathbb{E}\{p_1 - p_0\})$ values.

9.3.4 A Sensitivity Table

For reporting, a table tabulates the adjusted effect over a grid. Here for observed adjusted effect $\hat{\Delta}_{\text{obs}} = 2.0$ (cells show $\hat{\tau}_{\text{sens}} = \hat{\Delta}_{\text{obs}} - \delta(p_1 - p_0)$):

U-outcome effect δ	$p_1 - p_0 = 0.10$	$p_1 - p_0 = 0.30$	$p_1 - p_0 = 0.50$
weak ($\delta = 2$)	1.80	1.40	1.00
moderate ($\delta = 4$)	1.60	0.80	0.00
strong ($\delta = 8$)	1.20	−0.40	−2.00

The tipping point for the sign is the diagonal cell ($\delta = 4$, $p_1 - p_0 = 0.5$).

9.4 Three Canonical Sensitivity Models

The bias decomposition Equation 9.3 is a general identity. Three canonical *models* specialize it by imposing a structural restriction on either g_t or on the induced observed-data relationships. They are presented in order of increasing generality of the estimator they support.

9.4.1 Rosenbaum's Γ-Sensitivity Model

Definition: Γ-Sensitivity Model [@rosenbaum2002observational]

The Γ-sensitivity model at level $\Gamma \geq 1$ is the set of distributions $P(U \mid T, X)$ satisfying:

$$\frac{1}{\Gamma} \leq \frac{P(T = 1 \mid X, U = u)/P(T = 0 \mid X, U = u)}{P(T = 1 \mid X, U = u')/P(T = 0 \mid X, U = u')} \leq \Gamma. \tag{9.7}$$

$\Gamma = 1$ recovers conditional exchangeability given X alone. $\Gamma = 2$ states units with the same X may differ by up to a factor of 2 in treatment odds due to unmeasured U.

Theorem: Γ-Bounds on the Rank Statistic [@rosenbaum2002observational]

For a matched-pair design, the null distribution of the sign-rank statistic under the Γ-model is bounded by Binomial reference distributions with treatment probabilities $\Gamma/(1+\Gamma)$ and $1/(1+\Gamma)$.

The practical outcome is a tipping $\Gamma^\star$: the smallest Γ at which the test fails to reject. Small $\Gamma^\star$ (near 1) indicates a fragile conclusion; large $\Gamma^\star$ indicates a robust one.

9.4.2 The E-Value and the VanderWeele–Ding Bound

Ding and VanderWeele (2016) and VanderWeele and Ding (2017) proposed a sensitivity summary that requires no matched design and takes the form of a single closed-form number. It has become the most widely reported sensitivity summary in the applied literature.

Define two risk-ratio sensitivity parameters: $\text{RR}_{TU} = \max_x P(U = 1 \mid T = 1, X = x)/P(U = 1 \mid T = 0, X = x)$ (maximum RR of U with T), and $\text{RR}_{UY} = \max_{t,x} P(Y = 1 \mid T = t, X = x, U = 1)/P(Y = 1 \mid T = t, X = x, U = 0)$ (maximum RR of U with Y). The *bias factor* is:

$$B(\text{RR}_{TU}, \text{RR}_{UY}) = \frac{\text{RR}_{TU}\,\text{RR}_{UY}}{\text{RR}_{TU} + \text{RR}_{UY} - 1}. \tag{9.8}$$

Theorem: VanderWeele–Ding Bound

$\text{RR}^{\text{true}}_{TY|X} \geq \text{RR}^{\text{obs}}_{TY|X}/B(\text{RR}_{TU}, \text{RR}_{UY})$.
The observed risk ratio can be explained away only if both RR_{TU} and RR_{UY} are at least as large as

the **E-value**:

$$\mathrm{EV} = \mathrm{RR}^{\mathrm{obs}}_{TY|X} + \sqrt{\mathrm{RR}^{\mathrm{obs}}_{TY|X}\left(\mathrm{RR}^{\mathrm{obs}}_{TY|X} - 1\right)}. \tag{9.9}$$

Proof

Fix x and write $p_t = P(U{=}1 \mid T{=}t, X{=}x)$ and r_t for the outcome risk ratio of U at arm t. The ratio of observed to causal risk ratio is $[1+p_1(r_1-1)][1+p(r_0-1)]/\{[1+p(r_1-1)][1+p_0(r_0-1)]\}$. Maximizing over admissible (p_0, p_1, p, r_0, r_1) subject to $\max p_1/p_0 \le \mathrm{RR}_{TU}$ and $\max r_t \le \mathrm{RR}_{UY}$, the maximum is attained at $p_0 = 0$, $r_0 = 1$, giving the bound Equation 9.8.
For Equation 9.9: set $\mathrm{RR}_{TU} = \mathrm{RR}_{UY} = e$ and ask for the smallest e such that $B(e, e) = R$ (where $R = \mathrm{RR}^{\mathrm{obs}}_{TY|X}$). This gives $e^2/(2e-1) = R$, so $e = R + \sqrt{R(R-1)}$. □

Example: An E-Value Calculation

An observational study reports $\mathrm{RR}^{\mathrm{obs}}_{TY|X} = 2.5$ with 95% CI $[1.7, 3.7]$.
$\mathrm{EV} = 2.5 + \sqrt{2.5 \times 1.5} \approx 4.44$. To reduce the point estimate to the null, an unmeasured confounder must have both a RR association with treatment *and* an outcome-conditional RR association of at least 4.4.
$\mathrm{EV}_{\mathrm{CL}} = 1.7 + \sqrt{1.7 \times 0.7} \approx 2.79$. To reduce the lower CI to the null, the associations must each be at least 2.8.

Remark: Why a Single Number

The E-value collapses the two parameters to one by asking about the symmetric diagonal $\mathrm{RR}_{TU} = \mathrm{RR}_{UY} = e$. This is conservative: asymmetric configurations can also explain away the effect. The E-value is reported because it is a single interpretable scalar, not because it is the tightest possible summary.

9.4.3 The Marginal Sensitivity Model

Definition: Marginal Sensitivity Model [@tan2006distributional]

The *marginal sensitivity model* at level $\Lambda \ge 1$ bounds the odds ratio between the nominal propensity $\pi(x) = P(T = 1 \mid X = x)$ and the true propensity $\pi^{\mathrm{true}}(x, u) = P(T = 1 \mid X = x, U = u)$:

$$\frac{1}{\Lambda} \le \frac{\{1 - \pi^{\mathrm{true}}(X, U)\}\,\pi(X)}{\pi^{\mathrm{true}}(X, U)\,\{1 - \pi(X)\}} \le \Lambda \quad \text{a.s.} \tag{9.10}$$

$\Lambda = 1$ recovers no unmeasured confounding.

Theorem: MSM Bounds on the ATE [@zhao2019sensitivity]

Under the marginal sensitivity model at level Λ, the ATE satisfies $\tau_L(\Lambda) \le \tau \le \tau_U(\Lambda)$, where the sharp bounds are obtained by solving linear programs over the admissible weight perturbations induced by Equation 9.10.

Proof sketch. The true IPW weights equal the nominal weights multiplied by a perturbation factor $\phi_i \in [\Lambda^{-1}, \Lambda]$. Because the target is linear in ϕ, the extrema are attained at $\phi_i \in \{\Lambda^{-1}, \Lambda\}$. Zhao et al. (2019) give closed-form percentile expressions. □

Remark: MSM as the Estimation-Oriented Sensitivity Model

The MSM is the natural model to pair with IPW, AIPW, TMLE, and DML (Chapters 10–12), because it bounds the violation on the *weight* scale rather than on the unobservable $P(T \mid X, U)$. The sensitivity analysis reduces to a perturbation of the nominal weights by a factor in $[\Lambda^{-1}, \Lambda]$.

9.4.4 Comparing the Three Models

	Rosenbaum Γ	**E-value**	**MSM (Λ)**
Parameter bounds	odds ratio of treatment given X, U	pair of risk-ratio associations $\mathrm{RR}_{TU}, \mathrm{RR}_{UY}$	odds ratio of nominal to true propensity
Natural estimator	matched-pair and weighted rank tests	any relative-effect summary	IPW, AIPW
One-number summary	tipping $\Gamma^\star$	EV (symmetric diagonal)	bounds $[\tau_L, \tau_U]$ or tipping $\Lambda^\star$
Primary reference	Rosenbaum (2002)	VanderWeele and Ding (2017)	Tan (2006); Zhao et al. (2019)

Remark: Which Model to Report

Reporting more than one summary is common and useful. The E-value is almost always included because it is a single number that is widely understood; the Γ-value is reported in matched or rank-based analyses; and the MSM bounds are reported when the primary estimator is IPW or AIPW. The three are not substitutes: each bounds a different object.

9.5 Benchmarking Sensitivity Parameters

Every sensitivity model has parameters that are not identified by the data. A sensitivity curve of the form "at $\lambda = 0.2$ the effect is halved" is mathematically precise but scientifically empty until $\lambda = 0.2$ has been anchored to something concrete. *Benchmarking* gives sensitivity parameters an empirical referent by comparing them to the analogous quantities computed for the *observed* covariates (Cinelli and Hazlett 2020).

9.5.1 Benchmarking Against Observed Covariates

For the binary-confounder setup, compute for each observed covariate X_j: the coefficient on X_j in a regression of Y on $(T, X_1, \ldots)$ (the role of δ), and the difference in conditional means of X_j across treatment arms (the role of the imbalance). Plot these against the tipping contour.

For the E-value, for each X_j compute the bias factor:

$$B_j = \frac{\mathrm{RR}_{TX_j}\,\mathrm{RR}_{X_jY}}{\mathrm{RR}_{TX_j} + \mathrm{RR}_{X_jY} - 1}.$$

If $B_j <$ EV for every observed covariate, an unmeasured confounder inducing the observed effect would have to be at least as strong as some observed covariate. Cinelli and Hazlett (2020) develop partial-R^2 benchmarks for the linear regression setting.

Example: Benchmarked Reporting Language

The estimated effect of the job-training program on annual earnings is \$1,800 with 95% CI $[\$900, \$2,700]$. The E-value for the lower confidence limit is 2.3. Among the observed covariates, the largest bias factor is that of prior-year earnings, at $B = 2.0$; baseline education has $B = 1.5$. The conclusion is robust to an unmeasured confounder as strong as any observed covariate, but

could be overturned by a confounder approximately 15% stronger than the strongest observed one.

Sensitivity without Benchmarking

Reporting "the effect is robust up to $\Gamma = 2$" without a statement of whether $\Gamma = 2$ is plausible in the study at hand is the sensitivity-analysis analog of reporting a confidence interval without stating the confidence level: technically complete, but uninformative.

9.6 Partial Identification and Bounds

The sensitivity models of Section 9.4 each restrict the magnitude of a violation by a single parameter. A more radical approach imposes no parametric restriction at all — asking what the observed data can say about the causal parameter without any untestable identifying assumption. The answer is a set of values: the theory of *partial identification* (Manski 1990, 2003).

9.6.1 Point Identification vs. Partial Identification

Under *point identification*, $\psi = \Psi(P)$. Under *partial identification*, the assumptions pin down a set:

$$\psi \in \Psi_A(P) = \{\text{values of } \psi \text{ consistent with assumption set } A \text{ and distribution } P\}. \tag{9.11}$$

Definition: Sharp and Conservative Bounds

A bound $[\psi_L, \psi_U]$ is *sharp* under assumption set A if it equals $[\inf \Psi_A(P), \sup \Psi_A(P)]$. It is *conservative* (valid) if it contains the identified set but may be wider. Sharp bounds cannot be improved without adding assumptions; conservative bounds can sometimes be tightened by a more careful derivation.

9.6.2 Manski's No-Assumption Bound

Theorem: Manski's No-Assumption Bound [@manski1990nonparametric]

Suppose $Y \in [y_L, y_U]$ a.s. and $p = P(T = 1)$. Then $\tau \in [\tau_L, \tau_U]$ with:

$$\tau_L = [\mathbb{E}(Y \mid T = 1)\, p + y_L\,(1 - p)] - [\mathbb{E}(Y \mid T = 0)\,(1 - p) + y_U\, p], \tag{9.12}$$

$$\tau_U = [\mathbb{E}(Y \mid T = 1)\, p + y_U\,(1 - p)] - [\mathbb{E}(Y \mid T = 0)\,(1 - p) + y_L\, p]. \tag{9.13}$$

These bounds are **sharp**, and the width of the identified set is:

$$\tau_U - \tau_L = y_U - y_L. \tag{9.14}$$

Proof

By the law of total probability: $\mathbb{E}\{Y(1)\} = \mathbb{E}\{Y(1) \mid T = 1\}\, p + \mathbb{E}\{Y(1) \mid T = 0\}\,(1 - p)$ and $\mathbb{E}\{Y(0)\} = \mathbb{E}\{Y(0) \mid T = 1\}\, p + \mathbb{E}\{Y(0) \mid T = 0\}\,(1 - p)$. Consistency identifies $\mathbb{E}\{Y(1) \mid T = 1\} = \mathbb{E}(Y \mid T = 1)$ and $\mathbb{E}\{Y(0) \mid T = 0\} = \mathbb{E}(Y \mid T = 0)$. The unobserved counterfactual means $\mathbb{E}\{Y(1) \mid T = 0\}$ and $\mathbb{E}\{Y(0) \mid T = 1\}$ are restricted only to $[y_L, y_U]$. Taking the Minkowski difference of the resulting intervals for $\mathbb{E}\{Y(1)\}$ and $\mathbb{E}\{Y(0)\}$ gives Equation 9.12–Equation 9.13. Sharpness follows from the existence of degenerate potential-outcome distributions attaining each extreme. Width: $\tau_U - \tau_L = [y_U - y_L](1 - p) + [y_U - y_L]p = y_U - y_L$. □

Example: Manski Bound in a Binary Outcome

Suppose $Y \in \{0,1\}$, $P(T=1) = 0.4$, $\mathbb{E}(Y \mid T=1) = 0.6$, $\mathbb{E}(Y \mid T=0) = 0.3$. Then: $\tau_L = (0.6)(0.4) + (0)(0.6) - [(0.3)(0.6) + (1)(0.4)] = 0.24 - 0.58 = -0.34$; $\tau_U = (0.6)(0.4) + (1)(0.6) - [(0.3)(0.6) + (0)(0.4)] = 0.84 - 0.18 = 0.66$. The identified set $\tau \in [-0.34, 0.66]$ includes zero: the observed positive association of 0.3 is consistent with a causal effect anywhere from a meaningful harm to a substantial benefit.

The no-assumption bound is as wide as the support of the outcome — typically too wide to be informative. Productive partial identification therefore proceeds by adding *shape restrictions* that narrow the set.

Monotone treatment response ($Y(1) \geq Y(0)$ a.s.) tightens the bounds. **Monotone treatment selection** ($\mathbb{E}\{Y(t) \mid T=1\} \geq \mathbb{E}\{Y(t) \mid T=0\}$) is another common restriction. Combining these produces informative bounds even when neither alone suffices (Manski 2003).

Bounds as Sensitivity Analysis

Partial identification and sensitivity analysis are two views of the same underlying object. A sensitivity model with bounds $[\psi_L(\Lambda), \psi_U(\Lambda)]$ is a parametric partial-identification analysis; Manski's no-assumption bound is the limit when Λ permits arbitrary violations. Intermediate models (the Γ-model, the MSM) lie between these extremes, trading informativeness for robustness.

9.7 Sensitivity to Positivity and Overlap Violations

Chapter 6 introduced positivity as a structural requirement for the identification formulas underlying propensity-score methods. When $\pi(x_0) = 0$, the counterfactual mean $\mathbb{E}\{Y(1) \mid X = x_0\}$ is not identified, and the ATE may not be point-identified over the full covariate support. When positivity holds but weakly, IPW weights $1/\pi(X)$ generate extreme values that destabilize estimates.

9.7.1 Trimming as a Sensitivity Analysis

Definition: Trimmed ATE

$$\tau_\alpha = \mathbb{E}\{Y(1) - Y(0) \mid \alpha \leq \pi(X) \leq 1 - \alpha\}, \tag{9.15}$$

for $\alpha \in [0, 0.5)$. The choice $\alpha = 0$ recovers the full-population ATE.

Lemma: Trimmed Estimand Identity

Let $S_\alpha = \{X : \alpha \leq \pi(X) \leq 1 - \alpha\}$ and $q_\alpha = P(X \in S_\alpha) > 0$. Under ignorability:

$$\tau_\alpha = \frac{1}{q_\alpha} \mathbb{E}\{[\mu_1(X) - \mu_0(X)]\, \mathbf{1}\{X \in S_\alpha\}\}. \tag{9.16}$$

On the trimmed population, the IPW weights are bounded: $w(T, X) \leq 1/\alpha$ on S_α.

Proof. Conditional on $X \in S_\alpha$, positivity holds with slack α, so the back-door formula applies pointwise. Integrating and normalizing gives Equation 9.16. The weight bound follows directly from $\pi(X) \geq \alpha$ on S_α. □

τ_α and τ_0 are *different estimands*: trimming is a retargeting strategy, not a variance-reduction trick. A trimmed analysis answers "what is the causal effect on the subpopulation for which the treatment decision is not already essentially forced by covariates?"

Trimming Changes the Estimand

Reporting $\hat{\tau}_\alpha$ as a stabilized estimate of τ without acknowledging that the target population has changed is a common mistake. Whenever $\alpha > 0$, the estimand is Equation 9.15, not the full-population ATE. A proper sensitivity analysis reports $\hat{\tau}_\alpha$ and q_α together for a grid of $\alpha \in \{0, 0.01, 0.05, 0.10\}$.

9.8 Sensitivity to Invalid Instruments

Of the three IV assumptions, relevance is the most empirically diagnosable. Exogeneity and exclusion involve unobserved relationships and are not testable. A sensitivity analysis for IV concentrates on these two untestable assumptions.

Relevance is a statement about (Z, T) given X — observable. Exogeneity and exclusion are statements about (Z, Y) given X and the unobserved U — not observable. A complete robustness argument has two parts: (a) evaluation of first-stage strength and (b) sensitivity analysis for exogeneity and exclusion.

9.8.1 Direct-Effect Violation of Exclusion

Consider the scalar-IV model with a direct effect δ of Z on Y:

$$Y = \alpha + \beta T + \delta Z + \varepsilon, \qquad \mathbb{E}(\varepsilon \mid Z) = 0. \tag{9.17}$$

Theorem: IV Bias under Exclusion Violation

Under Equation 9.17 and standard IV regularity conditions:

$$\frac{\mathrm{Cov}(Y, Z)}{\mathrm{Cov}(T, Z)} = \beta + \delta \frac{\mathrm{Var}(Z)}{\mathrm{Cov}(T, Z)}. \tag{9.18}$$

Equivalently, the Wald estimand at posited δ is:

$$\beta(\delta) = \frac{\mathrm{Cov}(Y, Z) - \delta \,\mathrm{Var}(Z)}{\mathrm{Cov}(T, Z)}. \tag{9.19}$$

Proof. Take covariance of Equation 9.17 with Z: $\mathrm{Cov}(Y, Z) = \beta \,\mathrm{Cov}(T, Z) + \delta \,\mathrm{Var}(Z) + \mathrm{Cov}(\varepsilon, Z)$. By orthogonality, $\mathrm{Cov}(\varepsilon, Z) = 0$. Dividing and rearranging gives Equation 9.18 and Equation 9.19. □

Key pedagogical consequence: the bias is proportional to $\mathrm{Var}(Z)/\mathrm{Cov}(T, Z)$, the inverse of the first-stage slope. A *weak* first stage *amplifies* the bias from any exclusion violation.

Example: IV Sensitivity Curve

$\mathrm{Cov}(Y, Z) = 3$, $\mathrm{Cov}(T, Z) = 1.5$, $\mathrm{Var}(Z) = 1$. Wald estimate: $\hat{\beta}(0) = 2.0$. Applying Equation 9.19: $\hat{\beta}(\delta) = (3 - \delta)/1.5$. Values: $\hat{\beta}(0.5) = 1.67$, $\hat{\beta}(1.0) = 1.33$, $\hat{\beta}(2.0) = 0.67$, $\hat{\beta}(3.0) = 0$. The tipping point is $\delta^\star = 3$.

9.8.2 Connection to LATE and Monotonicity

Under heterogeneous treatment effects, sensitivity analysis can also address monotonicity violations, with the sensitivity parameter becoming the proportion of defiers (Angrist et al. 1996). The same conceptual framework — parameterize the violation, report a curve or bound — applies.

9.9 Sensitivity in Mediation Analysis

Sensitivity analysis is especially important in mediation analysis because the identifying assumptions are strictly stronger than those for total effects. Even in a randomized trial, Assumptions 3 and 4 of sequential ignorability cannot be guaranteed.

9.9.1 Residual-Correlation Sensitivity Parameter

Consider the linear mediation system $M = a_0 + aT + a_X^\top X + \varepsilon_M$, $Y = b_0 + \tau' T + bM + b_X^\top X + \varepsilon_Y$, with sensitivity parameter:

$$\rho = \mathrm{Corr}(\varepsilon_M, \varepsilon_Y \mid T, X). \tag{9.20}$$

If sequential ignorability holds, $\rho = 0$: disturbances are orthogonal because any shared source of variation has been conditioned out. A nonzero ρ represents unobserved M–Y confounding. Imai et al. (2010) derive the bias in the estimated NIE as a smooth function of ρ, giving a sensitivity curve $\rho \mapsto \widehat{\mathrm{NIE}}(\rho)$ with tipping point $\rho^\star$.

A well-reported mediation sensitivity analysis states: the point estimate of the indirect effect (with sampling CI), the tipping value $\rho^\star$, and a benchmark for what magnitude of ρ is plausible.

9.10 Sensitivity Analysis and Modern Estimators

Chapters 10–13 develop doubly robust estimation, orthogonal scores, cross-fitting, and semiparametrically efficient IV estimation. These tools make estimators more robust — but robust to a specific class of perturbations: *nuisance-model misspecification*. They are silent about identification.

What robust estimation does. Neyman orthogonality makes AIPW consistent if either the outcome model or propensity model is correctly specified (double robustness). With cross-fitting, nuisance estimators need only converge at rate $n^{-1/4}$ to yield $\sqrt{n}$-asymptotics. Both consequences address estimation under correctly maintained identifying assumptions.

What robust estimation does not do: Orthogonality and cross-fitting do not solve unmeasured confounding, positivity failure, IV invalidity, or mediation assumption failure.

> *Neyman orthogonality and cross-fitting protect the estimator against nuisance-estimation error. They do not protect against violations of causal identification assumptions.*

This is the reason sensitivity analysis belongs in Part II (identification) rather than Part III (estimation). Every method in Chapters 10–13 assumes identification and refines the estimation.

Applied Workflow (Recommended)

1. State the estimand (potential-outcome or do-notation).
2. State the identifying assumptions (back-door, IV, front-door, mediation, etc.).
3. Estimate the causal parameter (outcome regression, IPW, AIPW, DML, 2SLS, etc.).
4. Quantify sampling uncertainty (standard errors, confidence intervals).
5. Conduct sensitivity analysis for the least credible assumption (sensitivity curve, tipping point, or bounds, with benchmarking).
6. Report both the point estimate with sampling CI and the sensitivity summary, with plain-language interpretation of robustness.

9.11 Lab: A Tipping-Point Analysis for an Observational ATE

DGP. Let $X, U \overset{\text{i.i.d.}}{\sim} N(0,1)$ independently. Treatment: $P(T = 1 \mid X, U) = \mathrm{expit}(-0.3 + 0.6X + 1.0U)$. Outcome: $Y(t) = 1.5t + 0.8X + 2.0U + \varepsilon$, $\varepsilon \sim N(0,1)$. The analyst observes (Y, T, X) but not U. True ATE: $\tau = 1.5$.

Naive adjusted estimator. OLS coefficient on T in the regression of Y on $(1, T, X)$. Running 500 Monte Carlo replicates ($n = 2000$): mean of $\hat{\tau}_X$ is 3.172, SD is 0.099. Implied 95% CI $\approx [2.98, 3.37]$. The naive estimator is badly biased, but the CI is narrow and does not contain the truth. This is the identification-uncertainty failure mode in concrete form.

Sensitivity adjustment. Using the binary-U decomposition, bias is $B(\alpha_U, \delta) = \delta \cdot \mathrm{imbalance}(\alpha_U)$. Sensitivity-adjusted estimate $\hat{\tau}_{\text{sens}} = \hat{\tau}_X - B$ over a grid ($\hat{\tau}_X = 3.172$; the true configuration is $\alpha_U = 1.00, \delta = 2.0$):

Posited δ	$\alpha_U = 0.25$ (imb=0.232)	$\alpha_U = 0.50$ (imb=0.440)	$\alpha_U = 1.00$ (imb=0.787)	$\alpha_U = 1.50$ (imb=1.025)
0.5	3.056	2.952	2.779	2.660
1.0	2.940	2.732	2.385	2.147
2.0	2.708	2.292	**1.598**	1.123
3.0	2.477	1.851	0.811	0.098

The true configuration ($\alpha_U = 1.00, \delta = 2.0$) gives the sensitivity-adjusted value $1.598 \approx \tau = 1.5$ (checkmark).

Tipping points. $\delta^\star(\alpha_U) = \hat{\tau}_X / \text{imbalance}(\alpha_U)$. At $\alpha_U = 1.00$: $\delta^\star = 3.172/0.787 \approx 4.03$.

Benchmarking against observed X (outcome coefficient $\gamma_X = 0.8$): at $\alpha_U = 1.00$, the unmeasured confounder would need an outcome association about $4.03/0.8 \approx 5$ times stronger than X to zero out the estimate. A confounder as strong as X ($\delta = 0.8$, $\alpha_U = 1.0$) would adjust the estimate to ≈ 2.54 — still clearly positive.

Four lessons: (1) The naive CI $[2.98, 3.37]$ is narrow but misses the true ATE $\tau = 1.5$. Sampling uncertainty $\neq$ identification uncertainty. (2) Sensitivity adjustment at the correct configuration recovers the truth to within Monte Carlo error. (3) The tipping point depends jointly on both sensitivity parameters. (4) Benchmarking against observed X anchors the analysis: the conclusion "the effect is positive" is robust to unmeasured confounders up to about $5\gamma_X$ in outcome association.

9.12 Practical Reporting Guidelines

At minimum, a report should include: the point estimate $\hat{\psi}$ and its 95% CI; a sensitivity summary (curve, bound, E-value, or tipping point); a benchmarking statement; and a plain-language interpretation.

Misleading Phrases to Avoid

- "The result is causal because the estimate is statistically significant." Statistical significance addresses sampling uncertainty, not identification credibility.
- "The result is robust because we used machine learning." Flexible nuisance estimation improves the estimator; it does not validate the identifying assumptions.
- "Sensitivity analysis showed the result is not biased." Sensitivity analysis quantifies how much bias the identifying assumptions *could* produce if violated — it does not test for bias.
- "The E-value is X, therefore the effect is robust." An E-value is robust or not depending on whether X is plausible as an unmeasured-confounder strength. A large E-value without benchmarking is an unanchored statistic.

9.13 Chapter Summary

Symbol	Meaning
Δ_{obs}	Observed adjusted contrast
B	Confounding bias Equation 9.3
Γ	Rosenbaum odds-ratio bound
EV	E-value: $R + \sqrt{R(R-1)}$ Equation 9.9
Λ	MSM odds-ratio bound
τ_α	Trimmed ATE Equation 9.15
ρ	Residual correlation for mediation Equation 9.20
$\lambda^\star$	Tipping point for sign
$\lambda^\star_{\text{CI}}$	Tipping point for statistical significance

1. Sampling uncertainty, model misspecification, and identification uncertainty are distinct sources of error. A confidence interval addresses only the first, and identification uncertainty does not vanish merely by collecting more data from the same observational regime.

2. A sensitivity analysis introduces a parameter λ that quantifies the strength of a violation, with $\lambda = 0$ recovering the baseline assumption. The three reporting objects are the sensitivity curve, sensitivity bounds, and the tipping point.
3. The master bias decomposition $\Delta_{\text{obs}} = \tau + B$ underwrites sensitivity analysis for unmeasured confounding, specializing cleanly in the linear and binary-U cases.
4. Three canonical sensitivity models differ by what the sensitivity parameter bounds: Rosenbaum's Γ bounds a treatment-odds ratio; the E-value $\text{EV} = R + \sqrt{R(R-1)}$ gives a closed-form symmetric threshold; and the MSM bounds an odds ratio of nominal to true propensity.
5. Sensitivity parameters need benchmarking against observed covariates to be interpretable.
6. Partial identification reports an identified set rather than a single value. Manski's no-assumption bound is as wide as the support of the outcome; shape restrictions narrow it.
7. Positivity violations are an identification failure, not a variance problem. Trimming changes the estimand to a region of covariate overlap; it is a retargeting strategy, not a correction for the original population ATE.
8. Invalid instruments produce bias $\delta \cdot \text{Var}(Z)/\text{Cov}(T, Z)$ amplified by weak first stages. Weak-instrument diagnostics do not address exclusion violations.
9. Mediation sensitivity analysis proceeds by a residual-correlation parameter ρ that captures unmeasured M–Y confounding.
10. Modern estimation methods (AIPW, TMLE, DML, efficient GMM) address nuisance estimation, not identification. Neyman orthogonality protects the estimator; it does not protect the causal claim.

9.14 Problems

1. Sampling vs. identification uncertainty. Explain in your own words the difference between a narrow sampling confidence interval and a robust causal conclusion. Construct an example data-generating process, with numerical parameters, in which the CI for a back-door-adjusted ATE is very narrow (SD below 0.05) yet the true ATE lies outside it. Identify the identifying assumption that is violated and the magnitude of the violation.

2. Binary unmeasured confounder. Suppose $\hat{\Delta}_{\text{obs}} = 2.0$. Assume a binary unmeasured confounder U with constant outcome contrast $\delta = 4$ and imbalance $p_1 - p_0 = 0.3$.

(a) Use the Binary-Confounder Bias Lemma to compute B and the sensitivity-adjusted estimate $\hat{\tau}_{\text{sens}}$.
(b) What outcome contrast δ would zero out the estimate at the same imbalance level?
(c) What imbalance $p_1 - p_0$ would zero out the estimate at $\delta = 4$?

3. Tipping point in a linear bias model. For $\hat{\Delta}_{\text{obs}} = 1.5$ and bias model $B(\lambda) = 0.4\lambda$:

(a) Find the tipping point $\lambda^\star$ for the sign of the adjusted estimate.
(b) Suppose the sampling standard error is 0.3, independent of λ. Find the tipping point $\lambda^\star_{\text{CI}}$ for the lower confidence limit to reach zero (95% level). Compare to $\lambda^\star$.

4. Benchmarking. A researcher reports a sensitivity analysis in which the tipping point for the U-outcome effect is $\delta^\star = 3.0$. Observed covariate outcome coefficients are $(\hat{\gamma}_{\text{age}}, \hat{\gamma}_{\text{income}}, \hat{\gamma}_{\text{education}}) = (0.3, 1.8, 2.4)$. Identify the strongest benchmark. Is the result robust to an unmeasured confounder as strong as the strongest observed covariate? Write two sentences of interpretation suitable for an applied report.

5. E-value calculation. An observational study reports an adjusted risk ratio of $\text{RR}^{\text{obs}}_{TY|X} = 1.9$ with 95% CI $[1.3, 2.8]$.

(a) Compute the E-value for the point estimate using Equation 9.9.
(b) Compute the E-value for the lower confidence limit.
(c) Write one sentence of interpretation for each.

6. IV sensitivity curve. In a scalar-IV model with $\text{Cov}(Y, Z) = 3$, $\text{Cov}(T, Z) = 1.5$, $\text{Var}(Z) = 1$:

(a) Use Equation 9.19 to compute $\hat{\beta}(\delta)$ for $\delta \in \{0, 0.5, 1.0, 1.5, 2.0\}$.
(b) Find the tipping point $\delta^\star$.
(c) Repeat for the weaker-instrument case $\text{Cov}(T, Z) = 0.5$ (keeping other quantities fixed). Explain why the same δ produces a larger bias.

7. Manski bound calculation. Suppose $Y \in [0, 10]$, $P(T = 1) = 0.3$, $\mathbb{E}(Y \mid T = 1) = 6.5$, $\mathbb{E}(Y \mid T = 0) = 4.0$.

(a) Compute Manski's no-assumption bound using Equation 9.12–Equation 9.13.
(b) Compute the width and verify Equation 9.14.
(c) The observed association is 2.5. Does the no-assumption identified set include zero? Discuss what this implies about the informational content of the data alone.

8. Positivity sensitivity. Explain why the trimmed estimand τ_α of Equation 9.15 is generally not equal to the untrimmed ATE τ. In a dataset with $\pi(X)$ distributed uniformly on $[0, 1]$, what fraction of the population is retained at trimming levels $\alpha \in \{0.01, 0.05, 0.10, 0.20\}$? Under what scientific questions is τ_α preferable to τ as a target?

9. Mediation sensitivity. In a mediation study of a behavioral intervention (T) on depression (Y) through sleep quality (M), explain why randomization of T alone does not eliminate the need for a sensitivity analysis for the indirect effect. Draw the DAG that describes the residual concern and identify which arrow corresponds to a nonzero ρ in Equation 9.20. Give a plausible scientific story in which the residual correlation could be large.

10. Modern estimators and identification. A colleague argues that because DML and AIPW are "doubly robust and orthogonalized," they "automatically correct for unmeasured confounding provided the machine-learning models are good enough." Explain why this is incorrect, referring to the bias decomposition of the Bias Decomposition Theorem. Identify which term in that decomposition machine-learning methods can estimate consistently, and which term they cannot estimate at all.

Part III

Estimation

Chapter 10

Estimating Equations and Influence Functions

Learning Objectives

By the end of this chapter, students should be able to:

1. Explain why identification of a causal parameter does not automatically yield a statistically reliable estimator, and articulate the distinct questions that an estimation theory must address.
2. Define an estimating equation and verify the population moment condition for standard examples (sample mean, OLS, IPW).
3. State the definition of an asymptotically linear estimator and derive the influence function from a generic first-order expansion of an estimating equation.
4. Compute the influence function for simple causal estimators when nuisance functions are known, and interpret it as the infinitesimal contribution of a single observation.
5. Describe how estimation error in nuisance functions propagates into the target estimator via the Z-estimation stacked-equation framework, and explain why this is the central statistical challenge in causal inference.
6. Use the influence function to construct a consistent variance estimator and an approximate Wald confidence interval.
7. Define efficiency in the class of regular estimators and explain the role of the efficient influence function as the foundation for doubly robust estimation in Chapter 11.

10.1 Why Estimation Needs Its Own Theory

So far the focus has been on *identification*: under what assumptions can a causal parameter be written as a functional of the observed-data distribution? Identification does not by itself provide a statistically reliable estimator. Once a causal parameter is identified, several distinct questions remain:

- How should the estimator be constructed from the observed sample?
- How does estimation of nuisance functions affect the target estimator?
- What is the large-sample distribution of the estimator?
- How can we compute valid standard errors and confidence intervals?

These questions motivate a separate theory of *estimation and inference*. In causal inference this issue is especially important because identified parameters often depend on auxiliary, or *nuisance*, functions such as the outcome regression $\mu_t(x) = \mathbb{E}(Y \mid T{=}t,\, X{=}x)$ or the propensity score $\pi(x) = P(T{=}1 \mid X{=}x)$. Even when the causal parameter is identified, different estimation strategies may behave quite differently in terms of robustness, efficiency, and sensitivity to model misspecification.

The goal of this chapter is to introduce a general framework for estimation based on *estimating equations* and *influence functions*. This framework will serve as the foundation for the doubly robust and semiparametric methods developed in Chapter 11; see also Imbens and Rubin (2015) and Hernán and Robins (2020) for complementary treatments.

10.2 A Running Example: The ATE

Throughout this chapter we use the average treatment effect (ATE) as a running example:

$$\tau = \mathbb{E}\{Y(1) - Y(0)\}.$$

Under consistency, conditional exchangeability, and positivity, the ATE is identified by the back-door formula (Chapter 5):

$$\tau = \mathbb{E}[\mu_1(X) - \mu_0(X)], \qquad \mu_t(x) = \mathbb{E}(Y \mid T{=}t,\, X{=}x), \quad t \in \{0, 1\}.$$

This identity tells us what the target parameter is, but it does not uniquely determine how to estimate it. One may consider a *regression-based* estimator by fitting models for $\mu_1(x)$ and $\mu_0(x)$; a *weighting* estimator based on the propensity score $\pi(x)$; or an *augmented* estimator combining both. All of these may target the same causal parameter yet differ in their statistical properties. A main goal of this chapter is to develop a common language for describing and comparing such estimators.

10.3 Estimating Equations

A broad class of estimators can be defined as solutions to *estimating equations*. Let $O_1, \dots, O_n$ be i.i.d. observations from a distribution P, let θ denote a finite-dimensional target parameter, and write $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(O_i)$ for the empirical average.

Definition: Estimating Equation

An **estimating equation** for a parameter θ is an equation of the form $\mathbb{P}_n\{U(O;\theta)\} = 0$, where the **population moment condition** $\mathbb{E}\{U(O;\theta_0)\} = 0$ holds at the true parameter value θ_0. The function $U(O;\theta)$ is called the **estimating function**.

Many familiar estimators take this form.

Example: Three Estimating Equations

(i) Sample mean. Let $\theta = \mathbb{E}(Y)$. The sample mean $\hat{\theta} = \bar{Y}$ solves $\mathbb{P}_n(Y - \theta) = 0$. The estimating function is $U(O;\theta) = Y - \theta$.
(ii) Ordinary least squares. Let $O = (X, Y)$ and $\beta = [\mathbb{E}(XX^\top)]^{-1}\mathbb{E}(XY)$. OLS solves $\mathbb{P}_n[X\{Y - X^\top\beta\}] = 0$.
(iii) IPW estimating equation. Suppose $\pi(X)$ is known. Under consistency, conditional exchangeability, and positivity, $\tau = \mathbb{E}\{TY/\pi(X) - (1-T)Y/(1-\pi(X))\}$. Then τ solves:

$$\mathbb{E}\left[\frac{TY}{\pi(X)} - \frac{(1-T)Y}{1-\pi(X)} - \tau\right] = 0.$$

The right-hand side is an observed-data functional only after identification has been imposed; without those assumptions, knowing $\pi(X)$ alone does not turn the expression into a causal quantity.

The main advantage of the estimating-equation framework is that it provides a unified language for both estimator construction and asymptotic analysis.

10.4 From Estimating Equations to Asymptotic Linearity

Estimating equations are useful not only because they define estimators, but also because they often yield a convenient first-order expansion. Under suitable regularity conditions, an estimator solving an estimating equation can typically be approximated as $\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2} \sum_{i=1}^n \varphi(O_i) + o_p(1)$ for some mean-zero function $\varphi(O)$.

Definition: Asymptotic Linearity and Influence Function

An estimator $\hat{\theta}$ is **asymptotically linear** with **influence function** $\varphi(O)$ if

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(O_i) + o_p(1), \qquad \mathbb{E}\{\varphi(O)\} = 0. \tag{10.1}$$

This representation is fundamental because it immediately implies asymptotic normality by the CLT. When $\theta \in \mathbb{R}^p$:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0},\ \Sigma), \qquad \Sigma = \mathbb{E}[\varphi(O)\varphi(O)^\top].$$

In the scalar case this reduces to $\Sigma = \mathbb{E}[\varphi(O)^2]$.

Proposition: Generic First-Order Expansion

Under smoothness and regularity conditions, an estimator $\hat{\theta}$ solving $\mathbb{P}_n\{U(O;\theta)\} = 0$ admits a first-order expansion:

$$\sqrt{n}(\hat{\theta} - \theta_0) = -A^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U(O_i;\theta_0) + o_p(1), \qquad A = \mathbb{E}\left\{ \frac{\partial}{\partial \theta^\top} U(O;\theta_0) \right\}. \tag{10.2}$$

Hence the influence function is $\varphi(O) = -A^{-1}U(O;\theta_0)$.

Proof

A Taylor expansion of the sample moment condition around θ_0 gives:

$$0 = \mathbb{P}_n\{U(O;\hat{\theta})\} \approx \mathbb{P}_n\{U(O;\theta_0)\} + \mathbb{P}_n\left\{ \frac{\partial}{\partial \theta^\top} U(O;\theta_0) \right\} (\hat{\theta} - \theta_0).$$

By the LLN, $\mathbb{P}_n\{\partial_{\theta^\top} U(O;\theta_0)\} \to A$ in probability. Rearranging and multiplying by $\sqrt{n}$ yields the stated expansion. □

10.5 Influence Functions: Intuition

10.5.1 Statistical Functionals

Many parameters of interest in statistics and causal inference can be written as *functionals* of the underlying distribution.

Definition: Statistical Functional and Plug-In Estimator

A **statistical functional** is a map $\Psi : \mathcal{P} \to \mathbb{R}^p$ that assigns a parameter value $\psi = \Psi(P)$ to each distribution $P \in \mathcal{P}$. The **plug-in estimator** of ψ is $\hat{\psi} = \Psi(\mathbb{P}_n)$, obtained by replacing P with the empirical distribution.

Example: Common Statistical Functionals

(i) **Mean.** $\Psi(P) = \mathbb{E}_P(Y)$. Plug-in: $\bar{Y}$.
(ii) **Variance.** $\Psi(P) = \mathbb{E}_P(Y^2) - [\mathbb{E}_P(Y)]^2$. Plug-in: sample variance.
(iii) **Quantile.** $\Psi(P) = F_P^{-1}(\tau)$. Plug-in: sample τ-quantile.
(iv) **OLS regression coefficient.** $\Psi(P) = [\mathbb{E}_P(XX^\top)]^{-1}\mathbb{E}_P(XY)$. Plug-in: OLS estimator.
(v) **ATE.** $\Psi(P) = \mathbb{E}_P[\mu_1(X) - \mu_0(X)]$. Plug-in: average of estimated regression functions.

Remark: Linearity vs. Nonlinearity

A functional Ψ is **linear** if $\Psi((1-\epsilon)P + \epsilon Q) = (1-\epsilon)\Psi(P) + \epsilon\Psi(Q)$. The mean (i) is linear; the variance (ii), quantile (iii), OLS coefficient (iv), and ATE (v) are nonlinear. Linear functionals are straightforward to analyze because their plug-in estimators are sample averages. Nonlinear functionals require a local linearization, which is precisely what the influence function provides.

10.5.2 Influence Functions

Influence functions describe the first-order sensitivity of a statistical functional to small perturbations of the underlying distribution. Informally:

> *The influence function is the infinitesimal contribution of a single observation to the first-order behavior of the estimator.*

Definition: Influence Function (Informal)

A function $\varphi(O)$ is called an **influence function** for Ψ if it has mean zero under P and gives the first-order derivative of $\Psi(P)$ along regular parametric submodels:

$$\sqrt{n}(\hat{\psi} - \psi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(O_i) + o_p(1).$$

Remark: Formal Derivation via Pathwise Derivatives

The influence function can be defined formally through the *pathwise (Gateaux) derivative* of the functional $\Psi(P)$: if $P_\epsilon = (1-\epsilon)P + \epsilon\delta_O$ is the ϵ-mixture of P with a point mass at O, then:

$$\varphi(O) = \frac{d}{d\epsilon}\Psi(P_\epsilon)\Big|_{\epsilon=0}.$$

This derivative-based definition is the starting point for semiparametric efficiency theory. In this chapter it is enough to interpret $\varphi(O)$ as the object appearing in the asymptotic linear expansion.

Example: Influence Function of the OLS Slope via Pathwise Derivative

Consider $\beta = \Psi(P) = [\mathbb{E}(XX^\top)]^{-1}\mathbb{E}(XY)$.
Step 1. Let $P_\epsilon = (1-\epsilon)P + \epsilon\delta_O$. Then $\mathbb{E}_{P_\epsilon}(\tilde{X}\tilde{X}^\top) = (1-\epsilon)\mathbb{E}(XX^\top) + \epsilon XX^\top$ and similarly for $\mathbb{E}_{P_\epsilon}(\tilde{X}\tilde{Y})$.
Step 2. Differentiate with respect to ϵ at $\epsilon = 0$. Let $M(\epsilon) = (1-\epsilon)\mathbb{E}(XX^\top) + \epsilon XX^\top$ and $v(\epsilon) = (1-\epsilon)\mathbb{E}(XY) + \epsilon XY$. By differentiating $M(\epsilon)M(\epsilon)^{-1} = I$:

$$\frac{d}{d\epsilon}M(\epsilon)^{-1} = -M(\epsilon)^{-1}\dot{M}(\epsilon)\,M(\epsilon)^{-1}.$$

Evaluating at $\epsilon = 0$ and applying the product rule to $M^{-1}v$:

$$\varphi(O) = [\mathbb{E}(XX^\top)]^{-1}X(Y - X^\top\beta).$$

Verification. $\mathbb{E}\{\varphi(O)\} = [\mathbb{E}(XX^\top)]^{-1}\mathbb{E}[X(Y - X^\top\beta)] = [\mathbb{E}(XX^\top)]^{-1} \cdot 0 = 0$.
Connection to estimating equations. The OLS estimating function is $U(O;\beta) = X(Y - X^\top\beta)$, so $A = -\mathbb{E}(XX^\top)$ and Equation 10.2 gives $\varphi(O) = [\mathbb{E}(XX^\top)]^{-1}X(Y - X^\top\beta)$ — the same answer. The pathwise route is more fundamental; the estimating-equation route is more computationally direct.

10.6 Influence Functions for Simple Causal Estimators

We now illustrate the idea in causal settings when nuisance functions are *known*.

Known outcome regression. Suppose $\mu_1(\cdot)$ and $\mu_0(\cdot)$ are known. The natural plug-in estimator is $\hat{\tau} = \mathbb{P}_n\{\mu_1(X) - \mu_0(X)\}$. Its influence function is:

$$\varphi(O) = \mu_1(X) - \mu_0(X) - \tau.$$

Known propensity score. Suppose $\pi(X)$ is known. The IPW estimator $\hat{\tau} = \mathbb{P}_n\{TY/\pi(X) - (1-T)Y/(1-\pi(X))\}$ has influence function:

$$\varphi(O) = \frac{TY}{\pi(X)} - \frac{(1-T)Y}{1-\pi(X)} - \tau.$$

Remark: Nuisance Functions and the Influence Function

When nuisance functions are *known*, the influence function is simply the centered estimating function. This observation becomes much more consequential in the next section, where nuisance functions must be estimated. The key challenge is to understand how estimation error in $\hat{\mu}_t$ or $\hat{\pi}$ perturbs the first-order expansion.

Five Related Concepts: A Taxonomy

Several closely related objects share the symbol φ. Keeping them separate prevents confusion.
(i) Estimating function. A function $U(O;\theta)$ that defines an estimator through $\mathbb{P}_n\{U(O;\hat{\theta})\} = 0$. This is an *input* to the construction of $\hat{\theta}$.
(ii) Influence function of an estimator. A function $\varphi(O)$ such that $\sqrt{n}(\hat{\psi} - \psi_0) = n^{-1/2}\sum_i \varphi(O_i) + o_p(1)$. This is a *property of the estimator*. With known nuisance, φ is the centered estimating function; when nuisance is estimated, the two generally differ.
(iii) Influence function (canonical gradient) of a functional. A function $\varphi^*(O;P)$ that is the pathwise derivative of the parameter functional $\Psi(P)$ at P. This is a *property of the parameter and the model*, independent of any estimator.
(iv) Efficient influence function (EIF). Among all valid influence functions of regular estimators in the model, the unique one with smallest asymptotic variance.
(v) Estimated influence values. The numerical quantities $\hat{\varphi}_i = \hat{\varphi}(O_i)$, obtained by plugging in estimated nuisance functions. These are used to compute the sandwich variance $\widehat{V} = (n(n-1))^{-1}\sum_i(\hat{\varphi}_i - \bar{\varphi})^2$.
A regular estimator $\hat{\psi}$ is asymptotically linear with influence function φ^* (the canonical gradient) when nuisance estimators converge fast enough; the estimated influence values $\hat{\varphi}_i$ then serve as data-driven proxies for $\varphi^*(O_i)$ for variance estimation.

10.7 Z-Estimation with Nuisance Parameters

The examples in Section 10.6 treated nuisance functions as known. In practice, they must be estimated from the same data. When this happens, the randomness in the estimated nuisance introduces an additional term that the naive plug-in formula ignores. Z-estimation provides the framework that accounts jointly for estimation of both the target and nuisance parameters.

10.7.1 The Stacked Estimating Equation Framework

Suppose the target parameter $\psi \in \mathbb{R}^p$ depends on an unknown nuisance parameter $\alpha \in \mathbb{R}^q$. Both are estimated by solving a *stacked* system:

$$\mathbb{P}_n\{U_1(O;\psi,\alpha)\} = 0, \qquad \mathbb{P}_n\{U_2(O;\alpha)\} = 0. \tag{10.3}$$

Here U_1 defines the target estimator $\hat{\psi}$ given α, while U_2 defines the nuisance estimator $\hat{\alpha}$. The equations are solved simultaneously (or sequentially: first Equation 10.3 for $\hat{\alpha}$, then for $\hat{\psi}$).

Remark: Why Stack the Equations?

One might think it is enough to plug the estimated $\hat{\alpha}$ into the influence function derived under known α. This is incorrect in general: the randomness in $\hat{\alpha}$ contributes an additional term to the first-order expansion of $\hat{\psi}$ that is not captured by the plug-in influence function. The stacked framework keeps track of this extra term automatically.

10.7.2 The Z-Estimation Theorem

Write $\theta = (\psi^\top, \alpha^\top)^\top$ and let $U(O;\theta) = (U_1(O;\psi,\alpha)^\top, U_2(O;\alpha)^\top)^\top$ be the stacked estimating function.

Theorem: Z-Estimation Theorem (Stacked Equations)

Suppose: (i) $\mathbb{E}\{U(O;\theta_0)\} = 0$; (ii) $\hat{\theta} \xrightarrow{p} \theta_0$; (iii) the Jacobian $\mathcal{A} = \mathbb{E}\{\partial_{\theta^\top} U(O;\theta_0)\}$ is invertible; (iv) U is smooth enough in θ for a uniform LLN and CLT. Then:

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\mathcal{A}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U(O_i;\theta_0) + o_p(1).$$

Partitioning $\mathcal{A}$ conformably with (ψ,α) as $\mathcal{A} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$, the block-matrix inverse gives $\mathcal{A}^{-1} = \begin{pmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \\ 0 & A_{22}^{-1} \end{pmatrix}$, and the influence function of $\hat{\psi}$ is:

$$\varphi(O) = -A_{11}^{-1}\, U_1(O;\psi_0,\alpha_0) + A_{11}^{-1}A_{12}A_{22}^{-1}\, U_2(O;\alpha_0). \tag{10.4}$$

Proof

The argument mirrors Equation 10.2 applied to the full stacked system. Taylor-expand $\mathbb{P}_n\{U(O;\hat{\theta})\} = 0$ around θ_0:

$$0 \approx \mathbb{P}_n\{U(O;\theta_0)\} + \mathbb{P}_n\left\{\frac{\partial}{\partial\theta^\top} U(O;\theta_0)\right\}(\hat{\theta} - \theta_0).$$

By the LLN, the matrix of derivatives converges to $\mathcal{A}$. Rearranging and multiplying by $\sqrt{n}$ gives the stacked expansion. Extracting the first block via the block-matrix inverse formula gives Equation 10.4. □

Remark: Consistency Hypothesis

Hypothesis (ii) is a separate prerequisite, not a consequence of (i), (iii), and (iv). Standard arguments require identifiability of θ_0 via $\mathbb{E}\{U(O;\theta)\} = 0$ having a unique zero, together with uniform convergence of the sample moment. See Vaart (1998, Theorem 5.9) for a standard set of sufficient conditions.

Remark: Interpreting the Correction Term

Equation 10.4 decomposes the influence function into two parts. The first term $-A_{11}^{-1}U_1(O;\psi_0,\alpha_0)$ is exactly what we would obtain if α_0 were known. The second term $A_{11}^{-1}A_{12}A_{22}^{-1}U_2(O;\alpha_0)$ is the *nuisance correction* that accounts for the additional variability introduced by estimating α. If $A_{12} = 0$, the target estimating function U_1 does not depend on α at α_0, and the correction vanishes. This is the key condition exploited by *locally efficient* and *doubly robust* estimators in Chapter 11.

Remark: Just-Identified vs. Overidentified Moments

Equation 10.4 presupposes the system is **just identified**: U_1 has the same dimension as ψ and U_2 has the same dimension as α, so that A_{11} and A_{22} are square and invertible. When more moment conditions are available than parameters — the *overidentified* case from GMM — one minimizes $\mathbb{P}_n U^\top W \mathbb{P}_n U$ for some weight matrix W, and the influence function takes the standard GMM sandwich form (Hansen 1982). The block formula is the just-identified specialization.

10.7.3 Working Example: IPW with Estimated Propensity Score

We apply the Z-estimation theorem to the IPW estimator of the ATE τ when $\pi(X)$ is estimated by logistic regression. In Section 10.7.2 the generic target was ψ; here $\psi = \tau$.

Setup. Assume $\pi(X;\alpha) = \text{expit}(X^\top\alpha)$. The IPW estimating equation for τ is $U_1(O;\tau,\alpha) = TY/\pi(X;\alpha) - (1-T)Y/(1-\pi(X;\alpha)) - \tau$, and the logistic regression score is $U_2(O;\alpha) = X\{T - \pi(X;\alpha)\}$.

Computing the Jacobian blocks. Since $A_{11} = \mathbb{E}\{\partial_\tau U_1\} = -1$ and $A_{22} = \mathbb{E}\{\partial_{\alpha^\top} U_2\} = -\Sigma_\pi$ where $\Sigma_\pi = \mathbb{E}\{\pi(X;\alpha_0)(1-\pi(X;\alpha_0))XX^\top\}$. Using $\partial\pi/\partial\alpha^\top = \pi(1-\pi)X^\top$ and simplifying:

$$A_{12} = -\mathbb{E}\left[YX^\top\left\{\frac{T}{\pi(X;\alpha_0)} + \frac{1-T}{1-\pi(X;\alpha_0)} - 1\right\}\right].$$

Corrected influence function. Substituting into Equation 10.4 with $A_{11} = -1$ and $A_{22}^{-1} = -\Sigma_\pi^{-1}$ (the two negatives cancel):

$$\varphi(O) = U_1(O;\tau_0,\alpha_0) + A_{12}\,\Sigma_\pi^{-1}\,U_2(O;\alpha_0). \tag{10.5}$$

The first term is the naive IPW influence function from Section 10.6; the second is the nuisance correction removing the first-order contribution of $\hat{\alpha} - \alpha_0$.

Example: Variance Reduction from ML Estimation of the Propensity Score

We show that estimating α by maximum likelihood (ML) can only *reduce* the asymptotic variance of $\hat{\tau}_{\text{IPW}}$ relative to using the true α_0.
Write $\sigma_1^2 = \mathbb{E}[U_1(O;\tau_0,\alpha_0)^2]$. Expanding the variance of Equation 10.5:

$$\mathbb{E}[\varphi(O)^2] = \sigma_1^2 + 2A_{12}\Sigma_\pi^{-1}\mathbb{E}[U_1U_2] + A_{12}\Sigma_\pi^{-1}\mathbb{E}[U_2U_2^\top]\Sigma_\pi^{-1}A_{12}^\top. \tag{10.6}$$

Bartlett identity (ML score): $\mathbb{E}[U_2U_2^\top] = \Sigma_\pi$ (information equality for the correctly specified logistic model).
Cross-identity: Since $\mathbb{E}_\alpha[U_1(O;\tau_0,\alpha)] = 0$ for all α (under correct PS specification), differentiating through α: $A_{12} = -\mathbb{E}[U_1U_2^\top]$.
Substituting into Equation 10.6:

$$\mathbb{E}[\varphi(O)^2] = \sigma_1^2 - A_{12}\Sigma_\pi^{-1}A_{12}^\top = \sigma_1^2 - \mathbb{E}[U_1U_2^\top]\Sigma_\pi^{-1}\mathbb{E}[U_2U_1] \le \sigma_1^2,$$

since Σ_π is positive definite. Equality holds only when $\mathbb{E}[U_1U_2^\top] = 0$.
Scope. This variance-reduction conclusion (sometimes called the *estimated propensity score paradox*, Robins (1986)) requires: (i) correctly specified parametric propensity model; (ii) ML estimation so the Bartlett identity holds; (iii) strong overlap and moment conditions. When the parametric model is misspecified, when $\hat{\alpha}$ is computed by another method, or when $\pi(X)$ is estimated nonparametrically, the identity $A_{12} = -\mathbb{E}[U_1U_2^\top]$ need not hold and the conclusion fails.

The Central Challenge in Causal Estimation

The primary difficulty in practice is often not identifying the target parameter, but *correctly accounting for the effect of nuisance estimation* on the asymptotic distribution of the final estimator. Treating $\hat{\alpha}$ as if it were the true α_0 — computing standard errors from the naive IPW influence function rather than from Equation 10.4 — yields correct point estimates under a correctly specified PS model but **incorrect standard errors**. The Z-estimation framework resolves this variance-

accounting issue whenever the nuisance model is correctly specified and the nuisance estimator converges at rate $n^{-1/2}$. It does not, by itself, protect against model misspecification: if the parametric PS model is wrong, $\hat{\tau}$ is inconsistent regardless of which variance formula is used.

Remark: Machine Learning for Nuisance Estimation

The Z-estimation theorem as stated requires $\hat{\alpha}$ to converge at rate $n^{-1/2}$, which holds for finite-dimensional parametric models but fails for nonparametric or machine-learning estimators. When flexible methods are used for $\pi(x)$ or $\mu_t(x)$, the correction term in Equation 10.4 no longer has the simple form derived here, and a more careful analysis based on *sample splitting* or *cross-fitting* is needed. This is the starting point for the doubly robust and debiased machine-learning estimators in Chapter 11.

10.8 Variance Estimation and Confidence Intervals

Once an estimator admits the asymptotic linear representation, its asymptotic variance is $n^{-1}\mathrm{Var}\{\varphi(O)\}$. A natural estimator is obtained by replacing $\varphi(O_i)$ with estimated influence values $\hat{\varphi}_i$:

$$\hat{V} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\varphi}_i - \bar{\varphi})^2, \qquad \bar{\varphi} = \frac{1}{n} \sum_{i=1}^{n} \hat{\varphi}_i. \tag{10.7}$$

The asymptotically equivalent uncentered form $\hat{V} = n^{-2} \sum_i \hat{\varphi}_i^2$ is also commonly used; the two differ by $O_p(n^{-1})$. We adopt the centered form throughout this book. The approximate Wald confidence interval at level $1 - a$ is $\hat{\psi} \pm z_{1-a/2} \sqrt{\hat{V}}$.

Proposition: Variance from the Influence Function

If $\sqrt{n}(\hat{\psi} - \psi) = n^{-1/2} \sum_{i=1}^{n} \varphi(O_i) + o_p(1)$ with $\mathbb{E}\{\varphi(O)\} = 0$ and $\mathbb{E}\{\varphi(O)^2\} < \infty$, then:

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} N\left(0,\, \mathbb{E}[\varphi(O)^2]\right), \qquad \mathrm{Var}(\hat{\psi}) = \frac{1}{n}\, \mathbb{E}[\varphi(O)^2] + o(n^{-1}). \tag{10.8}$$

The proof follows immediately from the CLT applied to the i.i.d. sum. The influence function thus plays a dual role: it characterizes both the asymptotic distribution and the asymptotic variance.

Remark: Sandwich Variance Estimator

Equation 10.4 also provides the basis for the *sandwich variance estimator*. Replacing θ_0 by $\hat{\theta}$, the asymptotic variance of $\hat{\tau}$ is consistently estimated by Equation 10.7 with $\hat{\varphi}_i$ evaluated at $(\hat{\tau}, \hat{\alpha})$. This estimator automatically accounts for nuisance estimation uncertainty and is valid without any re-sampling. Standard errors from a logistic regression routine that ignores the link between $\hat{\alpha}$ and $\hat{\tau}$ will in general be **incorrect**.

Remark: Plug-In Influence Values in Practice

In practice, $\varphi(O_i)$ depends on unknown parameters and nuisance functions, so it is replaced by an estimated influence value $\hat{\varphi}_i$. This substitution is not innocuous: the plug-in variance formula is valid only when the estimator is asymptotically linear with influence function φ *and* the error from replacing unknown quantities by estimates is asymptotically negligible. For the plug-in estimators of Section 10.7, these conditions require nuisance estimators to converge fast enough that the plug-in error is $o_p(n^{-1/2})$. We return to this point in Chapter 11.

10.9 Toward Efficiency: Semiparametric Models and the EIF

Section 10.4 showed that the influence function determines both the asymptotic distribution and variance. Different regular estimators of the same parameter may therefore have different large-sample variances. This raises a natural question: among all regular estimators in a given model, what is the smallest achievable asymptotic variance? This section introduces the basic objects — semiparametric models, regular estimators, and the efficient influence function — as preparation for Chapter 11, where they are used to construct doubly robust and efficient estimators.

10.9.1 Semiparametric Models

Definition: Semiparametric Model

A **semiparametric model** is a statistical model $\mathcal{P}$ for the distribution P of the observed data in which: (i) the **parameter of interest** $\psi = \Psi(P) \in \mathbb{R}^p$ is finite-dimensional; and (ii) the remaining aspects of P — the **nuisance parameter** — are infinite-dimensional.

The canonical example in this course is the ATE. Under the identification assumptions of consistency, conditional exchangeability, and positivity, it equals the observed-data functional $\tau(P) = \mathbb{E}_P\{\mu_1(X) - \mu_0(X)\}$. We study estimation in the nonparametric model $\mathcal{P} = \{P : 0 < \pi(X) < 1 \text{ a.s.}\}$, where the nuisance consists of the entire joint distribution of (X, T, Y) subject only to positivity. No parametric form is assumed for $\mu_t(x)$ or $\pi(x)$.

Remark: Weak vs. Strong Overlap

The positivity condition $0 < \pi(X) < 1$ a.s. suffices for *identification* of τ, but not for stable regular root-n inference. Regular asymptotic theory typically requires **strong overlap**, $\epsilon \le \pi(X) \le 1 - \epsilon$ a.s. for some $\epsilon > 0$, together with appropriate moment conditions on Y. Under weak overlap the inverse-probability weights are unbounded in probability, and the asymptotic variance of IPW-type estimators may fail to be finite or fail to have a normal limiting distribution at the $\sqrt{n}$ rate.

Remark: Semiparametric vs. Parametric Efficiency

In a *fully parametric* model, the Cramér–Rao lower bound gives the minimum variance of any unbiased estimator of ψ. When the nuisance is infinite-dimensional, the classical bound no longer applies: the relevant lower bound must account for all infinitely many directions in which the likelihood can vary. The semiparametric efficiency bound is the analogue of the Cramér–Rao bound for this setting; see Bickel et al. (1993) and Tsiatis (2006) for comprehensive treatments.

10.9.2 Regular Estimators

The efficiency bound is meaningful only within the class of *regular* estimators, which informally are estimators whose asymptotic distribution is stable under small perturbations of the data-generating distribution.

Definition: Regular Estimator (Informal)

An estimator $\hat{\psi}$ of $\psi_0 = \Psi(P_0)$ is called **regular** at P_0 if, along every smooth parametric submodel $\{P_t : t \in \mathbb{R}\}$ with $P_0 = P_{t=0}$, the limiting distribution of $\sqrt{n}(\hat{\psi} - \psi_t)$ under $P_{1/\sqrt{n}}$ does not depend on the submodel or its direction.

Regularity rules out pathological estimators that exploit specific features of the data-generating mechanism in a non-uniform way (superefficient estimators). All estimators in this course — regression, IPW, and augmented variants — are regular. See Vaart (1998, Chs. 8, 25) and Tsiatis (2006, Ch. 4).

10.9.3 The Semiparametric Efficiency Bound and the EIF

Theorem: Semiparametric Convolution Theorem [@bickel1993]

In a semiparametric model $\mathcal{P}$, the asymptotic distribution of any regular estimator $\hat{\psi}$ of $\psi_0 \in \mathbb{R}^p$ satisfies:

$$\sqrt{n}(\hat{\psi} - \psi_0) \xrightarrow{d} N(0, V^*) * M$$

for some distribution M, where $*$ denotes convolution and $V^* = \mathbb{E}_P[\varphi^*(O)\,\varphi^*(O)^\top]$ is the **semiparametric efficiency bound**. The function $\varphi^*(O)$ with $\mathbb{E}[\varphi^*(O)] = 0$ and $\mathbb{E}[\varphi^*(O)\varphi^*(O)^\top] = V^*$ is called the **efficient influence function (EIF)**. An estimator achieves V^* if and only if it is asymptotically linear with influence function $\varphi^*(O)$.

Remark: Loewner Partial Order

For a scalar parameter ($p = 1$), V^* is a scalar and any regular estimator has asymptotic variance at least V^*. For vector-valued ψ, V^* is a $p \times p$ matrix and the comparison uses the **Loewner partial order**: any regular estimator has asymptotic covariance $V \succeq V^*$ (i.e., $V - V^*$ is positive semidefinite). Equivalently, every linear combination $c^\top \hat{\psi}$ has asymptotic variance at least $c^\top V^* c$ for every $c \in \mathbb{R}^p$.

Definition: Semiparametrically Efficient Estimator

A regular estimator $\hat{\psi}$ is **semiparametrically efficient** at $P_0 \in \mathcal{P}$ if $\sqrt{n}(\hat{\psi} - \psi_0) \xrightarrow{d} N(0, V^*)$, i.e., it achieves V^* with no additional convolution component. Equivalently, $\hat{\psi}$ is asymptotically linear with influence function $\varphi^*(O)$.

Remark: The EIF and the Parametric Score

It is tempting to describe the EIF as the "semiparametric score" of the model. The analogy is useful: $\varphi^*(O)$ plays the role in the semiparametric efficiency bound that the score plays in the Cramér–Rao bound. Strictly speaking, however, the score and the EIF live in different spaces. The score $\partial_\theta \log p(O; \theta_0)$ is an element of the model's *tangent space* (mean-zero functions reachable as derivatives of log-likelihoods). The EIF is the unique *canonical gradient*: the element of the tangent space that represents the derivative of the target functional Ψ via $\langle \varphi^*, S_\theta \rangle = \partial_\theta \Psi(P_\theta)|_{\theta=0}$. In a fully parametric model these coincide (scaled by the inverse Fisher information); in a semiparametric model they generally do not.

10.9.4 The Efficient Influence Function for the ATE

The EIF for the ATE

For the ATE $\tau = \mathbb{E}\{Y(1) - Y(0)\}$ in the nonparametric observed-data model under consistency, conditional exchangeability, and positivity, the efficient influence function is:

$$\varphi^*(O) = \frac{T\{Y - \mu_1(X)\}}{\pi(X)} - \frac{(1-T)\{Y - \mu_0(X)\}}{1 - \pi(X)} + \mu_1(X) - \mu_0(X) - \tau. \tag{10.9}$$

This expression has a natural decomposition: the first two terms are an IPW-style residual correction, and the last two are the regression estimator centered at τ. The EIF depends on *both* nuisance functions $\pi(X)$ and $\mu_t(X)$; an estimator that plugs in consistent estimates of both achieves the efficiency bound $V^* = \mathbb{E}[\varphi^*(O)^2]$.

Remark: Bridge to Chapter 11

Two important properties of Equation 11.11 will be central in Chapter 11:
(i) Double robustness. The estimator obtained by replacing $\pi(X)$ and $\mu_t(X)$ by estimates and averaging $\varphi^*(O_i)$ is consistent for τ if *either* $\hat{\pi}$ or $\hat{\mu}_t$ is consistent — not necessarily both. This is the *doubly robust* property.
(ii) Semiparametric efficiency. When both nuisance estimators are consistent and converge at suitable rates, the resulting estimator is asymptotically linear with influence function $\varphi^*(O)$ and therefore achieves the efficiency bound V^*.
The augmented IPW (AIPW) estimator, sometimes called the one-step or debiased estimator, is the principal tool for exploiting both properties simultaneously; it will be derived and analyzed in Chapter 11. The EIF Equation 11.11 becomes central there: the same object generates both doubly robust estimating equations and the semiparametric efficiency bound.

10.10 Chapter Summary

Object	Role
Estimating function $U(O;\theta)$	Defines the estimator via $\mathbb{P}_n\{U(O;\hat{\theta})\} = 0$
Influence function $\varphi(O)$	First-order term in expansion of $\hat{\psi} - \psi$; used for CLT and variance estimation
Asymptotic variance $\mathbb{E}[\varphi(O)^2]/n$	Governs precision; the target for efficiency comparisons
Efficient influence function $\varphi^*(O)$	Achieves the semiparametric efficiency bound; basis for doubly robust estimators (Chapter 11)
Sandwich variance estimator $\hat{V}$	Plug-in estimate of the asymptotic variance; valid under asymptotic linearity

1. **Identification is not estimation.** Identification provides a population formula, but not automatically a satisfactory estimator.
2. **Estimating equations.** Many estimators are defined as solutions to $\mathbb{P}_n\{U(O;\theta)\} = 0$, with the population moment condition identifying θ_0.
3. **Asymptotic linearity.** Estimating equations often yield a first-order expansion Equation 10.1, which immediately implies asymptotic normality via the CLT.
4. **Influence functions.** The function $\varphi(O)$ describes the first-order sensitivity of the estimator to a single observation. It is the key object for both asymptotic theory and variance estimation.
5. **Nuisance estimation.** In causal inference, nuisance functions must be estimated, and this estimation error propagates into the target estimator via the Z-estimation framework Equation 10.4. Controlling this propagation is the central statistical challenge.
6. **Variance from the influence function.** The asymptotic variance is $n^{-1}\mathbb{E}[\varphi(O)^2]$, consistently estimated by Equation 10.7.
7. **Efficiency.** Among regular estimators, the one with the smallest asymptotic variance is efficient. The efficient influence function characterizes this lower bound and guides estimator construction in Chapter 11.

10.11 Problems

1. Estimating equations and moment conditions.

(a) Let $O = (X, Y)$ and define $\theta = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$ (the coefficient in the population simple regression of Y on X). Write down an estimating equation for θ and verify the population moment condition holds at θ_0.
(b) Let $\theta = F^{-1}(0.5)$ be the population median. Propose an estimating function $U(O;\theta)$ and verify the population moment condition. (*Hint*: consider $U(O;\theta) = \mathbf{1}(Y \leq \theta) - 0.5$.)
(c) Show that the OLS estimator and the IPW estimator of the ATE are both special cases of the general M-estimator framework.

2. Asymptotic linearity and the CLT.

(a) Let $\widehat{\psi} = \bar{Y}$ be the sample mean of i.i.d. $Y_1, \dots, Y_n$ with $\mathbb{E}Y = \psi$ and $\mathrm{Var}(Y) = \sigma^2 < \infty$. Write down the influence function, state its asymptotic distribution, and give a consistent estimator of $\mathrm{Var}(\widehat{\psi})$.

(b) Suppose $\widehat{\psi}_1$ and $\widehat{\psi}_2$ are two asymptotically linear estimators of the same parameter with influence functions φ_1 and φ_2. Show that $\widehat{\psi}_\lambda = \lambda\widehat{\psi}_1 + (1-\lambda)\widehat{\psi}_2$ is also asymptotically linear for any fixed $\lambda \in \mathbb{R}$, and find its influence function.

(c) Use part (b) to derive the value of $\lambda \in \mathbb{R}$ that minimizes the asymptotic variance of $\widehat{\psi}_\lambda$. Express the answer in terms of $\sigma_1^2 = \mathbb{E}[\varphi_1^2]$, $\sigma_2^2 = \mathbb{E}[\varphi_2^2]$, and $\rho = \mathbb{E}[\varphi_1\varphi_2]$. Under what condition on $(\sigma_1^2, \sigma_2^2, \rho)$ does the optimal λ equal $1/2$?

3. Influence functions for causal estimators. Consider the ATE $\tau = \mathbb{E}\{Y(1) - Y(0)\}$ identified under consistency, conditional exchangeability, and positivity.

(a) Assume $\pi(X)$ and $\mu_t(X)$ are both known. Compute the influence functions of: (i) the regression estimator $\hat{\tau}_{\mathrm{reg}} = \mathbb{P}_n\{\mu_1(X) - \mu_0(X)\}$; (ii) the IPW estimator $\hat{\tau}_{\mathrm{IPW}} = \mathbb{P}_n\{TY/\pi(X) - (1-T)Y/(1-\pi(X))\}$.

(b) Under what conditions on the data-generating process will $\hat{\tau}_{\mathrm{reg}}$ have a smaller asymptotic variance than $\hat{\tau}_{\mathrm{IPW}}$?

(c) Verify that the EIF Equation 11.11 has mean zero under P by computing $\mathbb{E}[\varphi^*(O)]$.

4. Variance estimation.

(a) Let $\hat{\varphi}_i = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) - \hat{\tau}_{\mathrm{reg}}$ be the estimated influence values for the regression estimator. Write down the plug-in variance estimator $\hat{V}$ and simplify. Under what conditions is $\hat{V}$ consistent for $\mathrm{Var}(\hat{\tau}_{\mathrm{reg}})$?

(b) A researcher reports $\hat{\tau} = 2.4$ and $\hat{V} = 0.09$ based on $n = 400$ observations. Compute a 95% Wald confidence interval. Is the treatment effect statistically distinguishable from zero at the 5% level?

(c) Explain why simply plugging in $\hat{\mu}_t$ and $\hat{\pi}$ into the influence function formula may yield an inconsistent variance estimator when these nuisance estimators converge at a slower rate than $n^{-1/2}$.

5. Efficiency comparisons.

(a) Suppose φ_{reg} and φ_{IPW} denote the influence functions of the regression and IPW estimators. Without calculation, explain why neither estimator is always more efficient than the other.

(b) The semiparametric efficiency bound for the ATE is $\mathbb{E}[\varphi^*(O)^2]$, where φ^* is given in Equation 11.11. Show that $\mathbb{E}[\varphi^*(O)^2] \le \mathbb{E}[\varphi_{\mathrm{IPW}}(O)^2]$ by expanding the squared EIF. (*Hint*: use the law of iterated expectations to show that the cross-terms cancel.)

(c) Give one practical reason why an efficient estimator based on the EIF may not always be preferred over a simpler, less efficient estimator.

Chapter 11

Doubly Robust Estimation and Semiparametric Efficiency

Learning Objectives

By the end of this chapter, students should be able to:

1. Derive the AIPW estimator as a bias-corrected prediction estimator, and explain the role of each term.
2. Prove the double robustness property using the law of iterated expectations.
3. Describe the class of augmented IPW estimators indexed by arbitrary control functions $b_t(x)$, verify unbiasedness for any b_t, and identify the optimal choice that minimizes total variance.
4. Interpret the augmentation step as a Hilbert-space projection onto the orthogonal complement of the augmentation space, and derive the Pythagorean variance decomposition.
5. Carry out the projection argument in the causal inference setting, starting from the Horvitz–Thompson estimator.
6. State the semiparametric efficiency bound for the ATE and explain why the AIPW estimating function is the efficient influence function.
7. Derive two approaches to building double robustness directly into the outcome model: IPW-weighted regression and the augmented model with clever covariate; explain why including $\hat{\pi}_i^{-1}$ acts as a debiasing correction and connect this to TMLE.
8. Identify the product-rate condition as the key sufficient condition for nuisance estimation error to be asymptotically negligible, and construct a consistent variance estimator and Wald confidence interval.

11.1 Why Combine Outcome Regression and Weighting?

Under consistency, conditional exchangeability, and positivity, the ATE $\tau = \mathbb{E}\{Y(1) - Y(0)\}$ is identified by either of the following:

$$\tau = \mathbb{E}\{\mu_1(X) - \mu_0(X)\}, \qquad \text{or} \qquad \tau = \mathbb{E}\left\{\frac{TY}{\pi(X)} - \frac{(1-T)Y}{1-\pi(X)}\right\},$$

where $\mu_t(x) = \mathbb{E}(Y \mid T{=}t,\, X{=}x)$ and $\pi(x) = P(T{=}1 \mid X{=}x)$.

These formulas suggest two basic estimation strategies. The *outcome regression* (prediction) estimator averages $\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$ over the sample. The *IPW* estimator reweights observed outcomes using the propensity score. Each strategy has weaknesses: regression can be biased under misspecification, while IPW can be unstable when estimated propensity scores are close to 0 or 1 (Rosenbaum and Rubin 1983).

This chapter develops a third strategy: combine both models to construct an estimator that is consistent when *either* model is correct, and achieves the semiparametric efficiency bound when both are correct. We derive the same estimator in three complementary ways: as a bias-corrected prediction estimator

(Section 11.2), as the optimal member of a class of augmented estimators (Section 11.4 and Section 11.6), and as the efficient influence function (Section 11.7).

11.2 The Prediction Estimator and Its Bias

Work first with generic *working regression functions* $m_t(x)$, not necessarily equal to the true conditional means. The *prediction estimator* based on m_t is:

$$\hat{\tau}_{\text{pred}}(m) = \frac{1}{n}\sum_{i=1}^{n}\{m_1(X_i) - m_0(X_i)\}.$$

Define the population prediction estimand $\tau_{\text{pred}}(m) = \mathbb{E}\{m_1(X) - m_0(X)\}$. Under ignorability:

$$\tau_{\text{pred}}(m) - \tau = \mathbb{E}\big[(m_1(X) - \mu_1(X)) - (m_0(X) - \mu_0(X))\big]. \tag{11.1}$$

If a propensity score estimator $\hat{\pi}$ is available, the two bias terms in Equation 11.1 can be estimated from observed data. Defining residuals $e_i(t) = Y_i(t) - \hat{\mu}_t(X_i)$:

$$\widehat{\text{Bias}}(\hat{\tau}_{\text{pred}}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}\,e_i(1) + \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\hat{\pi}(X_i)}\,e_i(0).$$

The bias-corrected prediction estimator is $\hat{\tau}_{\text{AIPW}} = \hat{\tau}_{\text{pred}} - \widehat{\text{Bias}}(\hat{\tau}_{\text{pred}}) = \hat{\mu}_{1,\text{dr}} - \hat{\mu}_{0,\text{dr}}$, where:

$$\hat{\mu}_{1,\text{dr}} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}\{Y_i - \hat{\mu}_1(X_i)\} + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X_i), \tag{11.2}$$

$$\hat{\mu}_{0,\text{dr}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\hat{\pi}(X_i)}\{Y_i - \hat{\mu}_0(X_i)\} + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_0(X_i). \tag{11.3}$$

11.3 The Augmented IPW Estimator

Notation. From this point on, $\mu_t(x)$ inside estimating functions denotes a *generic working* outcome regression; the truth is written $\mu_t^*(x) = \mathbb{E}(Y \mid T{=}t, X{=}x)$. The propensity score $\pi(x)$ inside estimating functions denotes a working model; the truth is $\pi^*(x) = P(T{=}1 \mid X{=}x)$.

Collecting terms gives the estimating-equation form. The AIPW estimator solves $\mathbb{P}_n\{\phi(O;\tau,\hat{\eta})\} = 0$ where:

$$\phi(O;\tau,\eta) = \left[\frac{T}{\pi(X)}\{Y - \mu_1(X)\} + \mu_1(X)\right] - \left[\frac{1-T}{1-\pi(X)}\{Y - \mu_0(X)\} + \mu_0(X)\right] - \tau. \tag{11.4}$$

Solving explicitly:

Definition: AIPW Estimator

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{T_i}{\hat{\pi}(X_i)}\{Y_i - \hat{\mu}_1(X_i)\} - \frac{1-T_i}{1-\hat{\pi}(X_i)}\{Y_i - \hat{\mu}_0(X_i)\}\right]. \tag{11.5}$$

The **augmented inverse probability weighted (AIPW) estimator** equals the prediction estimator plus a weighted-residual bias correction.

Remark: Two Roles of the Augmentation Terms

The terms $T_i\{Y_i - \hat{\mu}_1(X_i)\}/\hat{\pi}(X_i)$ have two complementary roles. From the bias-correction perspective, they are IPW-weighted residuals that estimate the prediction error of the outcome regression. From an estimating-equation perspective, they make the estimating function orthogonal to nuisance

perturbations. The estimated bias term has zero expectation when the outcome model is correct; the bias-corrected estimator is unbiased when the propensity model is correct. Hence the name *doubly robust*.

Theorem: Double Robustness

Let $\phi(O;\tau,\eta)$ be the estimating function in Equation 11.4. Suppose consistency, conditional exchangeability, and positivity hold. Then $\mathbb{E}\{\phi(O;\tau,\eta)\} = 0$ if either:

1. $\mu_t(x) = \mu_t^*(x)$ for $t = 0, 1$, regardless of whether $\pi(x)$ is correctly specified; or
2. $\pi(x) = \pi^*(x)$, regardless of whether μ_0 and μ_1 are correctly specified.

Proof

Case 1: outcome regression correct. If $\mu_t(X) = \mathbb{E}(Y \mid T{=}t, X)$, then conditional on X:

$$\mathbb{E}\left[\frac{T}{\pi(X)}\{Y - \mu_1(X)\}\,\middle|\, X\right] = \frac{P(T{=}1 \mid X)}{\pi(X)}\,\mathbb{E}\{Y - \mu_1(X) \mid T{=}1, X\} = 0,$$

and similarly the untreated augmentation term vanishes. Therefore $\mathbb{E}\{\phi\} = \mathbb{E}\{\mu_1(X) - \mu_0(X)\} - \tau = 0$.

Case 2: propensity score correct. Suppose $\pi(X) = \pi^*(X)$ but μ_0, μ_1 are arbitrary. Compute the conditional expectation of the first bracket given X:

$$\mathbb{E}\left[\frac{T}{\pi(X)}\{Y - \mu_1(X)\} + \mu_1(X)\,\middle|\, X\right] = \frac{\mathbb{E}[TY \mid X]}{\pi(X)} - \mu_1(X)\,\frac{\mathbb{E}[T \mid X]}{\pi(X)} + \mu_1(X) = \mu_1^*(X),$$

using $\mathbb{E}[T \mid X] = \pi^*(X) = \pi(X)$ (so the μ_1 terms cancel) and $\mathbb{E}[TY \mid X] = \pi(X)\mu_1^*(X)$. By symmetry the control bracket has conditional expectation $\mu_0^*(X)$. Taking expectations over X and invoking $\tau = \mathbb{E}[\mu_1^*(X) - \mu_0^*(X)]$ gives $\mathbb{E}\{\phi\} = 0$. □

Remark: The Algebra of Case 2

The key cancellation is that the working outcome function $\mu_1(X)$ enters the bracket twice — once multiplied by $T/\pi(X)$ and once as a standalone term — and the two instances have equal and opposite conditional expectations because $\mathbb{E}[T/\pi(X) \mid X] = 1$ when π is correct. The bracket "forgets" μ_1 entirely and converges to $\mu_1^*(X)$. A wrong μ_t is subtracted out by the correct π.

Double robustness guarantees consistency if *either* model is correct; it does *not* protect against misspecification of both models simultaneously.

11.4 A Class of Augmented Estimators

Remark: Three Senses of "Optimal" in This Chapter

(i) **Class-optimal** control function b_t^* minimizing variance within the parametric family (Section 11.4); (ii) **Projection-optimal** estimator $\hat{\theta}_{\text{opt}}$ minimizing variance after orthogonal augmentation correction (Section 11.5); (iii) **Semiparametrically efficient** estimator attaining the information lower bound (Section 11.7). The first two agree in the ATE setting, and Section 11.7 shows they also agree with the third.

For any square-integrable functions $b_1(x)$ and $b_0(x)$, define:

$$\hat{\mu}_{1,b} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi(X_i)}\,Y_i - \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi(X_i)} - 1\right\} b_1(X_i), \tag{11.6}$$

$$\hat{\mu}_{0,b} = \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\pi(X_i)}\,Y_i - \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1-T_i}{1-\pi(X_i)} - 1\right\} b_0(X_i), \tag{11.7}$$

and let $\hat{\tau}_b = \hat{\mu}_{1,b} - \hat{\mu}_{0,b}$. The standard AIPW estimator is the special case $b_t = \mu_t$ (since $TY/\pi - (T/\pi - 1)\mu_1 = T(Y-\mu_1)/\pi + \mu_1$).

Theorem: Unbiasedness and Variance of the AIPW Class

Under conditional exchangeability, SUTVA, and positivity, with the true propensity score, for any square-integrable b_0, b_1:

1. **Unbiasedness:** $\mathbb{E}(\hat{\tau}_b) = \tau$.
2. **Total variance:**

$$\mathrm{Var}(\hat{\tau}_b) = \frac{1}{n}\mathbb{E}\left[\left(\frac{1}{\pi(X)} - 1\right)\{Y(1) - b_1(X)\}^2 + 2\{Y(1) - b_1(X)\}\{Y(0) - b_0(X)\} + \left(\frac{1}{1-\pi(X)} - 1\right)\{Y(0) - b_0(\tag{11.8}$$

Proof

Let $\xi = \frac{T}{\pi(X)}\{Y - b_1(X)\} + b_1(X) - \frac{1-T}{1-\pi(X)}\{Y - b_0(X)\} - b_0(X)$, so $\hat{\tau}_b = n^{-1}\sum_i \xi_i$. By SUTVA, $\xi = \frac{T}{\pi(X)}\{Y(1) - b_1(X)\} + b_1(X) - \frac{1-T}{1-\pi(X)}\{Y(0) - b_0(X)\} - b_0(X)$.

Part (i). Since $\mathbb{E}[T/\pi(X) \mid X] = 1$, $\mathbb{E}[(T/\pi(X) - 1)b_1(X)] = 0$. Under conditional exchangeability, $\mathbb{E}[(T/\pi(X))Y(1) \mid X] = \mu_1^*(X)$. Hence $\mathbb{E}[\hat{\mu}_{1,b}] = \mu_1$ and $\mathbb{E}(\hat{\tau}_b) = \tau$.

Part (ii). Apply the law of total variance conditioning on $(X, Y(1), Y(0))$. The conditional mean is $\mathbb{E}[\xi \mid X, Y(1), Y(0)] = Y(1) - Y(0)$. Since ξ is linear in T with $\mathrm{Var}(T \mid X) = \pi(X)(1-\pi(X))$:

$$\mathrm{Var}(\xi \mid X, Y(1), Y(0)) = \left(\frac{1}{\pi(X)} - 1\right)\{Y(1) - b_1(X)\}^2 + 2\{Y(1) - b_1(X)\}\{Y(0) - b_0(X)\} + \left(\frac{1}{1-\pi(X)} - 1\right)\{Y(0) - b_0(X$$

Adding the between-group term $\mathrm{Var}(Y(1) - Y(0))$ gives Equation 11.8. □

Theorem: Optimal Control Functions

The control functions $b_1^*(X) = \mathbb{E}\{Y(1) \mid X\}$ and $b_0^*(X) = \mathbb{E}\{Y(0) \mid X\}$ minimize the total variance in Equation 11.8. Under conditional exchangeability and consistency, $b_t^*(X) = \mu_t^*(X) = \mathbb{E}(Y \mid T{=}t, X)$.

Proof

Write $u(X) = \mu_1^*(X) - b_1(X)$ and $v(X) = \mu_0^*(X) - b_0(X)$, and decompose $Y(t) - b_t(X) = d_t(X) + \varepsilon_t$ where $d_1 = u$, $d_0 = v$, $\mathbb{E}[\varepsilon_t \mid X] = 0$. Cross terms between (u, v) and $(\varepsilon_1, \varepsilon_0)$ vanish by iterated expectations, leaving a b_t-dependent term:

$$Q(b_0, b_1) = \mathbb{E}\left[\left(\frac{1}{\pi} - 1\right)u^2 + 2uv + \left(\frac{1}{1-\pi} - 1\right)v^2\right].$$

Writing the integrand as a perfect square:

$$\frac{1-\pi}{\pi}u^2 + 2uv + \frac{\pi}{1-\pi}v^2 = \left(\sqrt{\frac{1-\pi}{\pi}}\,u + \sqrt{\frac{\pi}{1-\pi}}\,v\right)^2 \geq 0,$$

with equality at $u \equiv 0$, $v \equiv 0$, i.e., $b_t^* = \mu_t^*$. □

Remark: Non-Uniqueness of the Joint Minimizer

The perfect square vanishes whenever $(1-\pi(X))u(X) + \pi(X)v(X) = 0$ a.s., so $Q = 0$ on a one-dimensional family of pairs (b_0, b_1). The canonical choice $b_t^* = \mu_t^*$ is distinguished by being the arm-wise optimum, separately minimizing $\mathrm{Var}(\hat{\mu}_{t,b})$ for each t. The arm-by-arm projection in Section 11.6 makes this uniqueness explicit.

11.5 The Projection Interpretation

The optimality of AIPW has an elegant interpretation in terms of projections in a Hilbert space of estimators. The collection of mean-zero square-integrable random variables under inner product $\langle X, Y\rangle = \mathbb{E}(XY) = \mathrm{Cov}(X, Y)$ is a genuine Hilbert space, and the results below are instances of the L^2 projection theorem.

Let $\hat{\theta}_0$ be an unbiased estimator of θ. Define the *augmentation space* Λ as a closed linear subspace of mean-zero square-integrable random variables computable from the observed data without knowledge of θ. For any $\hat{b} \in \Lambda$ the estimator $\hat{\theta}_b = \hat{\theta}_0 - \hat{b}$ remains unbiased.

Theorem: Optimal Projection

The optimal correction is $\hat{b}^* = \Pi(\hat{\theta}_0 \mid \Lambda)$, the L^2 projection of $\hat{\theta}_0$ onto Λ, characterized by: (1) $\hat{b}^* \in \Lambda$; (2) $\mathrm{Cov}(\hat{\theta}_0 - \hat{b}^*, \hat{b}) = 0$ for all $\hat{b} \in \Lambda$. The optimal estimator is:

$$\hat{\theta}_{\text{opt}} = \hat{\theta}_0 - \hat{b}^* = \Pi(\hat{\theta}_0 \mid \Lambda^{\perp}),$$

and satisfies the **Pythagorean identity**:

$$\mathrm{Var}(\hat{\theta}_0) = \mathrm{Var}(\hat{\theta}_{\text{opt}}) + \mathrm{Var}(\hat{b}^*). \tag{11.9}$$

Proof Sketch

Λ is a closed linear subspace, so the L^2 projection theorem gives unique $\hat{b}^* \in \Lambda$ satisfying the orthogonality condition. Any other choice $\hat{b} \in \Lambda$ yields:

$$\mathrm{Var}(\hat{\theta}_0 - \hat{b}) = \mathrm{Var}(\hat{\theta}_{\text{opt}}) + \mathrm{Var}(\hat{b}^* - \hat{b}) \geq \mathrm{Var}(\hat{\theta}_{\text{opt}}),$$

using $\mathrm{Cov}(\hat{\theta}_{\text{opt}}, \hat{b}^* - \hat{b}) = 0$ (since $\hat{b}^* - \hat{b} \in \Lambda$ and $\hat{\theta}_{\text{opt}} \perp \Lambda$). Specializing to $\hat{b} = 0$ gives Equation 11.9. □

The Pythagorean identity shows the variance of the initial estimator decomposes orthogonally. Tsiatis (2006) calls Λ the *augmentation space* because its elements are the corrections that reduce variance.

Remark: Finite-Sample Analogy for Semiparametric Projection

This theorem is a finite-sample $L^2(P)$ analogy for the projection arguments in semiparametric efficiency theory. In the full theory, the projection is at the level of influence functions and tangent spaces. Here we project a realized estimator $\hat{\theta}_0$ onto the orthogonal complement of an augmentation space Λ of observable corrections. Section 11.6 shows that the finite-sample projection recovers AIPW, and Section 11.7 identifies the AIPW estimating function with the efficient influence function from the tangent-space projection.

11.6 Projection in the Causal Inference Setting

We now specialize the projection framework to the ATE, assuming $\pi(X)$ is known. Decompose the Horvitz–Thompson estimator as $\hat{\tau}_{\text{HT}} = \hat{\mu}_{1,\text{HT}} - \hat{\mu}_{0,\text{HT}}$ where:

$$\hat{\mu}_{1,\text{HT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\pi_i}, \qquad \hat{\mu}_{0,\text{HT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\pi_i}.$$

Define arm-specific augmentation spaces:

$$\Lambda_1 = \left\{ n^{-1}\sum_{i=1}^{n}\left(\frac{T_i}{\pi_i} - 1\right) b_1(X_i) : b_1 \in \mathcal{L}^2 \right\}, \quad \Lambda_0 = \left\{ n^{-1}\sum_{i=1}^{n}\left(\frac{1-T_i}{1-\pi_i} - 1\right) b_0(X_i) : b_0 \in \mathcal{L}^2 \right\}. \tag{11.10}$$

Every element of Λ_1 (resp. Λ_0) has expectation zero, so augmenting leaves each arm-mean estimator unbiased. The combined augmentation space is $\Lambda = \Lambda_1 + \Lambda_0$.

Theorem: Arm-wise Projections onto the Augmentation Spaces

The optimal corrections are:

$$\Pi(\hat{\mu}_{1,\mathrm{HT}} \mid \Lambda_1) = n^{-1} \sum_{i=1}^{n} \left(\frac{T_i}{\pi_i} - 1 \right) b_1^*(X_i), \quad b_1^*(x) = \mathbb{E}\{Y(1) \mid x\},$$

$$\Pi(\hat{\mu}_{0,\mathrm{HT}} \mid \Lambda_0) = n^{-1} \sum_{i=1}^{n} \left(\frac{1 - T_i}{1 - \pi_i} - 1 \right) b_0^*(X_i), \quad b_0^*(x) = \mathbb{E}\{Y(0) \mid x\}.$$

Proof

We verify the treated arm. Denote the residual $R_1 = \hat{\mu}_{1,\mathrm{HT}} - n^{-1} \sum_i (T_i/\pi_i - 1) b_1^*(X_i) = n^{-1} \sum_i r_{1i}$ where $r_{1i} = (T_i/\pi_i)\{Y_i - b_1^*(X_i)\} + b_1^*(X_i)$.
By the Optimal Projection Theorem, it suffices to show $\mathrm{Cov}(R_1, \hat{b}) = 0$ for every $\hat{b} = n^{-1} \sum_j (T_j/\pi_j - 1) b_1(X_j) \in \Lambda_1$. Using independence across units, $\mathrm{Cov}(R_1, \hat{b}) = n^{-1} \mathbb{E}[r_{1i}\, b_i]$. Condition on $(X_i, Y_i(1), Y_i(0))$; the only randomness is $T_i \mid X_i$. Since $T_i^2 = T_i$, $\mathbb{E}[T_i/\pi_i \cdot (T_i/\pi_i - 1) \mid X_i] = 1/\pi_i - 1$. Taking outer expectation and using $\mathbb{E}\{Y_i(1) - b_1^*(X_i) \mid X_i\} = 0$, the contribution vanishes. The constant term $b_1^*(X_i)$ in r_{1i} also contributes zero because $\mathbb{E}[(T_i/\pi_i - 1) \mid X_i] = 0$. □

The optimal estimators are:

$$\hat{\mu}_{1,\mathrm{opt}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{T_i}{\pi_i} \{Y_i - \mu_1^*(X_i)\} + \mu_1^*(X_i) \right], \quad \hat{\mu}_{0,\mathrm{opt}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1 - T_i}{1 - \pi_i} \{Y_i - \mu_0^*(X_i)\} + \mu_0^*(X_i) \right].$$

Their difference is precisely the AIPW estimator Equation 11.5 with $b_t^*(X) = \mu_t^*(X)$, confirming the Optimal Control Functions theorem by a different route.

Remark: What the Arm-wise Decomposition Adds

The joint optimization in the Unbiasedness and Variance theorem has a non-unique minimizer (Remark on Non-Uniqueness). The arm-wise projection has a *unique* minimizer $b_t^*(x) = \mathbb{E}\{Y(t) \mid X{=}x\}$, and the pair (μ_0^*, μ_1^*) also attains the joint minimum. The arm-wise projection therefore selects the canonical representative, and the resulting estimator coincides with AIPW.

11.7 The Efficient Influence Function and Semiparametric Efficiency

This section states the semiparametric efficiency result and interprets it in light of the estimator development above. A complete proof requires explicit characterization of the nuisance tangent space; see Tsiatis (2006) for a rigorous development.

Under the nonparametric model for $O = (X, T, Y)$, the efficient influence function for the ATE is:

$$\varphi_{\mathrm{eff}}(O) = \frac{T}{\pi(X)} \{Y - \mu_1(X)\} - \frac{1 - T}{1 - \pi(X)} \{Y - \mu_0(X)\} + \mu_1(X) - \mu_0(X) - \tau. \tag{11.11}$$

Comparing Equation 11.11 with Equation 11.4, the AIPW estimating function *is* the efficient influence function $\varphi^*(O)$ from Chapter 10. The augmentation space Λ in Equation 11.10 is the finite-sample analogue of the nuisance tangent space of the nonparametric model, and projection onto $\Lambda^\perp$ is the finite-sample counterpart of the semiparametric operation that removes nuisance tangent directions.

Theorem: Semiparametric Efficiency Bound for the ATE

Under the nonparametric model for $O = (X, T, Y)$, every regular asymptotically linear estimator $\hat{\tau}$ of the ATE satisfies:

$$\text{Avar}(\sqrt{n}(\hat{\tau} - \tau)) \;\geq\; \mathbb{E}\{\varphi_{\text{eff}}(O)^2\},$$

and the bound is attained by estimators whose influence function equals φ_{eff} in Equation 11.11.

Remark: EIF as the Semiparametric Score

The efficient influence function plays the same role in semiparametric theory that the score plays in parametric models: its variance gives the information lower bound, analogous to the Cramér–Rao bound. Strictly speaking, φ_{eff} is not itself a score (the score belongs to the tangent space); it is the *canonical gradient* — the Riesz representer of the pathwise derivative of τ along regular submodels (Bickel et al. 1993).
The three derivations in this chapter — bias correction (Section 11.2), optimal augmentation (Section 11.4), and semiparametric identification — converge on the same estimating function. This convergence is the deepest explanation of why AIPW occupies a central place in the causal inference toolkit.

11.8 Doubly Robust Regression: Weighted and Augmented Approaches

The AIPW estimator achieves double robustness by adding an explicit bias-correction term. An equally important question is how to build double robustness *directly into the outcome model fit*, so that the prediction estimator is itself doubly robust without a separate augmentation step. The unifying concept is the *internal bias calibration* (IBC) condition.

11.8.1 The Internal Bias Calibration Conditions

Let $\hat{\mu}_t(x)$ denote any fitted outcome model and $\hat{\pi}_i = \hat{\pi}(X_i)$. The prediction estimator requires no augmentation if the IPW-weighted residuals vanish:

$$\sum_{i=1}^{n} \frac{T_i}{\hat{\pi}_i}\{Y_i - \hat{\mu}_1(X_i)\} = 0, \qquad \sum_{i=1}^{n} \frac{1 - T_i}{1 - \hat{\pi}_i}\{Y_i - \hat{\mu}_0(X_i)\} = 0. \tag{11.12}$$

We call Equation 11.12 the *internal bias calibration* (IBC) conditions (Firth and Bennett 1998). When both IBC conditions hold, the augmentation terms in Equation 11.5 are zero by construction, so $\hat{\tau}_{\text{pred}} = \hat{\tau}_{\text{AIPW}}$ and the prediction estimator is itself doubly robust.

11.8.2 Weighted Regression Approach

Suppose the outcome model for arm t is parameterized as $\mu_t(X; \theta_t)$ with a constant term. Use *IPW-weighted* least squares:

$$\sum_{i=1}^{n} \frac{T_i}{\hat{\pi}_i}\{Y_i - \mu_1(X_i; \theta_1)\}^2. \tag{11.13}$$

The normal equation of Equation 11.13 with respect to the intercept component is $\sum_i \frac{T_i}{\hat{\pi}_i}\{Y_i - \mu_1(X_i; \hat{\theta}_1)\} = 0$, which is exactly IBC condition Equation 11.12. Hence IPW-weighted fitted values automatically satisfy IBC for any model containing a constant (Robins et al. 1994; Bang and Robins 2005). The key point is not that weighted regression creates a fundamentally different doubly robust estimator; rather, it constructs fitted values for which the prediction estimator algebraically equals the AIPW estimator.

11.8.3 Augmented Model Approach and the Clever Covariate

Let $\hat{\mu}_1^{(0)}(X_i)$ be any initial fit. Augment it by including the *clever covariate* $\hat{\pi}_i^{-1}$ (Laan and Rubin 2006; Laan and Rose 2011) and run OLS of Y_i on $\hat{\mu}_1^{(0)}(X_i)$ and $\hat{\pi}_i^{-1}$ among treated units:

$$Y_i = \alpha_1 + \beta_1 \hat{\mu}_1^{(0)}(X_i) + \gamma_1 \hat{\pi}_i^{-1} + e_i(1), \qquad T_i = 1. \tag{11.14}$$

The fitted outcome model is $\hat{\mu}_1(X_i) = \hat{\alpha}_1 + \hat{\beta}_1 \hat{\mu}_1^{(0)}(X_i) + \hat{\gamma}_1 \hat{\pi}_i^{-1}$.

The normal equation for $\hat{\gamma}_1$ (the coefficient on $\hat{\pi}_i^{-1}$) is $\sum_i \frac{T_i}{\hat{\pi}_i}\{Y_i - \hat{\mu}_1(X_i)\} = 0$, exactly IBC condition Equation 11.12. Since this sum equals zero, the prediction estimator is:

$$\hat{\mu}_1^{\text{pred}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\hat{\pi}_i}\{Y_i - \hat{\mu}_1(X_i)\},$$

which is the AIPW representation. The covariate $\hat{\pi}_i^{-1}$ is "clever" because it is chosen so that the fitted regression satisfies the same score equation appearing in the AIPW bias correction.

> Remark: Behavior under Correct Outcome Model
>
> If $\hat{\mu}_1^{(0)}$ is correctly specified, then $\hat{\alpha}_1 \xrightarrow{p} 0$, $\hat{\beta}_1 \xrightarrow{p} 1$, and $\hat{\gamma}_1 \xrightarrow{p} 0$: the augmented model reduces asymptotically to $\hat{\mu}_1^{(0)}(X_i)$, and including the additional covariates carries no asymptotic efficiency cost.

> Remark: Connection to TMLE
>
> The augmented model approach captures the same targeting idea that underlies *targeted minimum loss-based estimation* (TMLE) (Laan and Rubin 2006; Laan and Rose 2011). The standard TMLE targeting step uses the fixed-offset form (coefficient on $\hat{\mu}_1^{(0)}$ fixed at 1), the minimal fluctuation sufficient to satisfy the IBC condition. The model Equation 11.14 goes further by freely estimating $\hat{\beta}_1$, providing additional recalibration of the initial fit beyond canonical TMLE.

11.9 Lab: Simulation Study

This lab compares four estimators of the ATE: $\hat{\tau}_{\text{HT}}$, $\hat{\tau}_{\text{AIPW}}$, the fixed-offset augmented-model estimator $\hat{\tau}_{\text{aug}}$, and the improved augmented-model estimator $\hat{\tau}_{\text{aug}^+}$ of Equation 11.14. A 2×2 design over nuisance-model correctness demonstrates double robustness directly. The lab also reports empirical 95% Wald coverage, illustrating the distinction between double-robust consistency and valid asymptotic inference.

DGP. $n = 1000$ i.i.d. observations. Draw $X_i \sim N(0,1)$, $T_i \mid X_i \sim \text{Bernoulli}(\pi^*(X_i))$ with $\pi^*(x) = \text{expit}\{0.2x + 0.2(x^2 - 1)\}$. Potential outcomes: $Y_i(1) = 1 + X_i + 0.5X_i^2 + \varepsilon_i(1)$, $Y_i(0) = X_i + 0.5X_i^2 + \varepsilon_i(0)$, $\varepsilon_i(t) \sim N(0,1)$ i.i.d., giving true ATE $\tau = 1$. The X^2 term enters both the true OR and true PS; omitting it yields four scenarios.

Scenarios.

Scenario	OR fit	PS fit
S1	correct: $Y \sim (1, X, X^2)$	correct: $T \sim (1, X, X^2)$
S2	correct: $Y \sim (1, X, X^2)$	misspecified: $T \sim (1, X)$
S3	misspecified: $Y \sim (1, X)$	correct: $T \sim (1, X, X^2)$
S4	misspecified: $Y \sim (1, X)$	misspecified: $T \sim (1, X)$

The fitted $\hat{\pi}$ is clipped to $[10^{-3}, 1 - 10^{-3}]$.

Results ($B = 2000$ replications, `set.seed(2025)`). Bias, Var, MSE $\times 10^{-3}$; Cov = empirical 95% Wald coverage.

Scenario	Metric	$\hat{\tau}_{\text{HT}}$	$\hat{\tau}_{\text{AIPW}}$	$\hat{\tau}_{\text{aug}}$	$\hat{\tau}_{\text{aug}^+}$
S1: OR , PS	Bias	0.55	−0.44	−0.45	−0.45
	Var	6.82	4.25	4.25	4.30
	MSE	6.82	4.25	4.25	4.30
	Cov (%)	99.9	95.0	94.9	94.5
S2: OR , PS	Bias	188.56	0.62	0.62	0.62
	Var	6.32	4.33	4.33	4.33
	MSE	41.87	4.32	4.32	4.32
	Cov (%)	67.4	94.2	94.2	94.2
S3: OR , PS	Bias	1.46	2.02	2.25	−25.26
	Var	5.74	4.71	4.41	8.02
	MSE	5.74	4.71	4.41	8.66
	Cov (%)	99.9	98.4	98.2	93.2
S4: OR , PS	Bias	188.42	188.77	188.62	−5.42
	Var	6.28	6.30	6.29	4.26
	MSE	41.78	41.93	41.87	4.29
	Cov (%)	67.5	33.1	33.1	94.7

S1 (both correct). All four estimators are approximately unbiased. $\hat{\tau}_{\text{HT}}$ has about 60% higher MSE because it discards the outcome model. The three augmented estimators are essentially equivalent, confirming $(\hat{\alpha}_t, \hat{\beta}_t, \hat{\gamma}_t) \to (0, 1, 0)$ when the OR is correct.

S2 (OR correct, PS misspecified). HT is badly biased (MSE $\approx 42 \times 10^{-3}$, coverage 67.4%). The three augmented estimators are indistinguishable: when the OR is correct, they converge to the same limit regardless of PS specification. **First direct demonstration of double robustness.**

S3 (OR misspecified, PS correct). HT, AIPW, and $\hat{\tau}_{\text{aug}}$ remain consistent through the correct PS. $\hat{\tau}_{\text{aug}}$ achieves the lowest MSE (4.41×10^{-3}). $\hat{\tau}_{\text{aug}^+}$ shows a finite-sample bias (-25×10^{-3}) from high-leverage clever-covariate values. **Second double-robustness demonstration.**

S4 (both misspecified). Double robustness provides no guarantee. HT, AIPW, and $\hat{\tau}_{\text{aug}}$ are all badly biased; Wald coverage collapses to $\approx 33\%$ for AIPW and aug. The apparent "recovery" of $\hat{\tau}_{\text{aug}^+}$ (bias -5×10^{-3}, coverage 94.7%) is a DGP-specific artifact and does **not** indicate triple robustness.

Takeaway. Comparing S2–S3 against S4 is the crisp illustration of **?@thm-dr**: AIPW is consistent whenever at least one nuisance model is correct, and only S4 breaks the guarantee.

11.10 Asymptotic Inference with Estimated Nuisance Functions

Under suitable regularity conditions:

$$\sqrt{n}(\hat{\tau}_{\text{AIPW}} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{\text{eff}}(O_i) + o_p(1) \xrightarrow{d} N(0,\ \mathbb{E}\{\varphi_{\text{eff}}(O)^2\}).$$

The $o_p(1)$ remainder captures nuisance estimation error. Neyman orthogonality implies this error enters only through a second-order remainder. A sufficient condition for the remainder to be negligible is the **product-rate condition**:

$$\|\hat{\pi} - \pi\| \cdot \|\hat{\mu}_t - \mu_t\| = o_p(n^{-1/2}), \qquad t = 0, 1, \tag{11.15}$$

where $\|\cdot\|$ denotes the $L_2(P)$ norm. A symmetric sufficient condition: each nuisance estimator converges at rate $o_p(n^{-1/4})$. This is much weaker than the parametric rate $n^{-1/2}$ required of each individually.

Remark: Neyman Orthogonality

The first-order insensitivity of the AIPW estimating function to nuisance perturbations is an instance of *Neyman orthogonality*, a property shared by a broad class of semiparametric estimators. Chapter 12 exploits this property systematically through cross-fitting, which eliminates the need for additional empirical-process conditions on the nuisance estimators.

11.10.1 Variance Estimation and Confidence Intervals

Because $\hat{\tau}_{\mathrm{AIPW}}$ is asymptotically linear with influence function φ_{eff}, its asymptotic variance is estimated by the empirical variance of the plug-in influence values:

$$\hat{\varphi}_i = \frac{T_i}{\hat{\pi}(X_i)}\{Y_i - \hat{\mu}_1(X_i)\} - \frac{1-T_i}{1-\hat{\pi}(X_i)}\{Y_i - \hat{\mu}_0(X_i)\} + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) - \hat{\tau}_{\mathrm{AIPW}}, \tag{11.16}$$

$$\hat{V} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\varphi}_i - \bar{\varphi})^2, \qquad \bar{\varphi} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varphi}_i. \tag{11.17}$$

The Wald confidence interval is $\hat{\tau}_{\mathrm{AIPW}} \pm z_{1-\alpha/2}\sqrt{\hat{V}}$.

Remark: Double Robustness versus Efficient Inference

It is important not to conflate two distinct results. *Double-robust consistency* (**?@thm-dr**) requires only that one nuisance model be consistent, together with overlap and standard LLN regularity. It does *not* require the product-rate condition Equation 12.12.

Efficient asymptotic inference — the $\sqrt{n}$-normal expansion and the Wald interval based on $\hat{V}$ — is a strictly stronger requirement, needing *both* nuisance functions to be estimated consistently at rates satisfying Equation 12.12. Plugging estimated nuisance functions into Equation 11.17 estimates the asymptotic variance of the *efficient* influence function only when the asymptotic linear representation with φ_{eff} holds; that representation does *not* follow from double robustness alone.

Suppose the propensity score is correct but the outcome regression is misspecified: $\hat{\pi} \to \pi^*$ but $\hat{\mu}_t \to \mu_t^\dagger \neq \mu_t^*$. AIPW remains consistent for τ (**?@thm-dr**, Case 2): the AIPW moment evaluated at the limits $(\pi^*, \mu_0^\dagger, \mu_1^\dagger)$ is

$$\varphi^\dagger(O) = \frac{T}{\pi^*(X)}\{Y - \mu_1^\dagger(X)\} - \frac{1-T}{1-\pi^*(X)}\{Y - \mu_0^\dagger(X)\} + \mu_1^\dagger(X) - \mu_0^\dagger(X) - \tau, \tag{11.18}$$

with $\mathbb{E}\{\varphi^\dagger(O)\} = 0$.

It is tempting to conclude that $\varphi^\dagger$ is then the influence function of $\hat{\tau}_{\mathrm{AIPW}}$ and that the plug-in variance Equation 11.17 consistently estimates $\mathbb{E}\{\varphi^\dagger(O)^2\}$. This conclusion is correct only when π^* is *known*, not estimated. When π is estimated and $\mu^\dagger \neq \mu^*$, the AIPW moment is no longer Neyman-orthogonal in the π-direction at $(\pi^*, \mu^\dagger)$. The Gateaux derivative of the population moment with respect to π at this point, along a perturbation $h(X)$, evaluates to

$$-\mathbb{E}\left[h(X)\left\{\frac{\mu_1^*(X) - \mu_1^\dagger(X)}{\pi^*(X)} + \frac{\mu_0^*(X) - \mu_0^\dagger(X)}{1-\pi^*(X)}\right\}\right],$$

which is generally nonzero whenever $\mu_t^\dagger \neq \mu_t^*$. The first-order error from estimating π therefore enters the asymptotic linear expansion of $\sqrt{n}(\hat{\tau}_{\mathrm{AIPW}} - \tau)$, acquiring a contribution beyond $n^{-1/2}\sum_i \varphi^\dagger(O_i)$. Consequently, the plug-in variance Equation 11.17 is not generally valid under one-correct-model misspecification, and the naive Wald interval can have incorrect coverage. Valid asymptotic inference in this regime requires one of the following: (i) treating $(\hat{\pi}, \hat{\mu}, \hat{\tau}_{\mathrm{AIPW}})$ jointly as the solution of a stacked estimating-equation system and using the corresponding sandwich variance, which absorbs the nuisance-estimation contribution; (ii) sample splitting or cross-fitting together with conditions ensuring the nuisance-estimation contribution is asymptotically negligible (Chapter 12); or (iii) verifying that the misspecified limit coincides with the truth, which restores orthogonality. The plug-in variance Equation 11.17 is justified without qualification only when $\hat{\tau}_{\mathrm{AIPW}}$ admits an asymptotic linear representation with φ_{eff} — a representation that holds at (π^*, μ^*) but generally fails at $(\pi^*, \mu^\dagger)$ once π^* is replaced by its estimator.

Extreme Propensity Scores

When $\hat{\pi}(X_i)$ is near 0 or 1, the IPW weights become large and can destabilize the estimator. Practical remedies: overlap diagnostics, truncation of extreme propensity scores, or restricting the target population to a subgroup with adequate overlap. The augmented model approach of Section 11.8 provides a complementary strategy by folding the propensity score into the outcome-model fit.

Looking ahead: when plug-in fails. For finite-dimensional parametric nuisance models, the product-rate condition holds automatically at the parametric rate $n^{-1/2}$. The situation changes with flexible machine-learning methods: nuisance convergence rates may be slower than $n^{-1/4}$, and machine-learning function classes are typically not Donsker, so the empirical-process remainder need not vanish when the same data are used for both nuisance estimation and score evaluation.

Chapter 12 addresses both difficulties by (a) formalizing Neyman orthogonality and exhibiting the AIPW score as an orthogonal score, and (b) decoupling nuisance estimation from score evaluation via *cross-fitting*. The resulting *double/debiased machine learning* (DML) estimator (Chernozhukov et al. 2018) is a direct extension of the AIPW development here.

11.11 Comparison of Regression, IPW, and AIPW

All consistency statements below are conditional on the causal identification assumptions of Chapter 5: consistency, conditional exchangeability, and positivity.

Estimator	**Uses μ_t**	**Uses π**	**Consistent if**	**Main weakness**
Regression (prediction)	Yes	No	outcome model correct	Sensitive to OR misspecification
IPW (Horvitz–Thompson)	No	Yes	propensity model correct	Unstable under weak overlap
AIPW	Yes	Yes	either model correct	Requires estimating both nuisances and careful inference

When overlap is weak, the IBC-based approaches of Section 11.8 can improve numerical stability by folding the propensity score into the outcome-model fit. They do not, however, eliminate the fundamental information loss where covariate support is lacking: in such cases the practical recommendation is to change the target estimand (trimming, overlap weighting, or restriction to a subgroup) rather than to expect any algebraic refinement of AIPW to repair the problem.

11.12 Chapter Summary

Symbol	**Meaning**
τ	ATE $= \mathbb{E}\{Y(1) - Y(0)\}$
$\mu_t(x)$	Outcome regression $\mathbb{E}(Y \mid T{=}t, X{=}x)$
$\pi(x)$	Propensity score $P(T{=}1 \mid X{=}x)$
$b_t(x)$	Control function; optimal $b_t^*(x) = \mu_t^*(x)$
$\hat{\tau}_{\text{AIPW}}$	AIPW estimator Equation 11.5
Λ	Augmentation space Equation 11.10
$\varphi_{\text{eff}}(O)$	Efficient influence function Equation 11.11
IBC	Internal bias calibration conditions Equation 11.12
$\hat{\pi}_i^{-1}$	Clever covariate; its normal equation enforces IBC

1. The prediction estimator is biased when the outcome model is misspecified; the bias can be estimated with the propensity score and subtracted to yield the AIPW estimator.

2. The AIPW estimator has two key properties: *double robustness* (consistent when either model is correct) and *class-optimal variance* (the choice $b_t^* = \mu_t^*$ minimizes variance when both nuisances are correctly specified).
3. The same class-specific optimum follows from a Hilbert-space projection: the optimal estimator is the projection of the HT estimator onto $\Lambda^\perp$, and the Pythagorean identity governs the variance reduction.
4. The AIPW estimating function is the efficient influence function for the ATE; its variance gives the semiparametric efficiency bound. The bias-correction, optimal-augmentation, and semiparametric-efficiency derivations all converge on the same object.
5. Double robustness can be enforced within the outcome-regression fitting step, by IPW-weighted regression or by an augmented regression model including the inverse propensity score as a clever covariate (TMLE connection).
6. The simulation (Section 12.10) demonstrates double robustness across a 2×2 design: AIPW remains consistent whenever at least one nuisance model is correct (S1–S3) and loses the guarantee in S4. Wald coverage is approximately nominal in S1–S3 and collapses in S4.
7. For asymptotic inference, the product-rate condition Equation 12.12 is sufficient for root-n inference. In modern machine-learning settings, cross-fitting makes these conditions more plausible; Chapter 12 develops this approach.

11.13 Problems

1. Bias of the prediction estimator.

(a) Verify the bias formula Equation 11.1 by writing $\tau_{\text{pred}}(m) - \tau$ as a difference of working-model errors and simplifying.
(b) Under what condition on the propensity score model does $\widehat{\text{Bias}}(\hat{\tau}_{\text{pred}})$ have zero expectation? Provide a careful argument using iterated expectations.
(c) Construct a simple example (binary X, binary T, binary Y) in which the estimated bias is nonzero and the outcome model is misspecified, but the AIPW estimator is still consistent.

2. The AIPW class and optimal control functions.

(a) Verify that $\hat{\tau}_b$ with $b_t = \mu_t$ equals the AIPW estimator Equation 11.5.
(b) Using Equation 11.8, show that $b_t^*(x) = \mathbb{E}\{Y(t) \mid x\}$ minimizes conditional variance by completing the square in b_t.
(c) Suppose $\pi(x) = 1/2$ for all x. Simplify Equation 11.8 and interpret the result.

3. Double robustness.

(a) Verify Case 1 of **?@thm-dr**: with $\mu_t = \mu_t^*$ and arbitrary π, show $\mathbb{E}\{\phi\} = 0$.
(b) Verify Case 2: with $\pi = \pi^*$ and arbitrary μ_t, show $\mathbb{E}\{\phi\} = 0$.
(c) Provide a counterexample showing $\mathbb{E}\{\phi\} \neq 0$ when both models are misspecified.

4. Projection and the Pythagorean identity.

(a) Verify $\text{Cov}(\hat{\theta}_{\text{opt}}, \hat{b}^*) = 0$ from the definition of $\hat{b}^*$.
(b) Deduce Equation 11.9 from the decomposition $\hat{\theta}_0 = \hat{\theta}_{\text{opt}} + \hat{b}^*$.
(c) Explain why $\hat{\theta}_{\text{opt}}$ is still unbiased for θ even though $\hat{b}^* \in \Lambda$ is subtracted.

5. Semiparametric efficiency.

(a) Show $\mathbb{E}\{\varphi_{\text{eff}}(O)^2\} \leq \mathbb{E}\{\varphi_{\text{IPW}}(O)^2\}$ by writing $\varphi_{\text{IPW}} = \varphi_{\text{eff}} + (\varphi_{\text{IPW}} - \varphi_{\text{eff}})$ and showing the cross term vanishes.
(b) Interpret the efficiency gain $\mathbb{E}\{\varphi_{\text{IPW}}^2\} - \mathbb{E}\{\varphi_{\text{eff}}^2\}$ in terms of the variance of the outcome regression.
(c) Give one reason why an efficient estimator based on φ_{eff} may still be disfavored in practice relative to a simpler, less efficient alternative.

6. Augmented model and the clever covariate.

(a) Let $\hat{\mu}_1^{(0)}(X_i)$ be any initial outcome model fit. Show that the normal equation for $\hat{\gamma}_1$ in the augmented model Equation 11.14 is exactly the IBC condition Equation 11.12, and conclude that the prediction estimator can always be written in AIPW form.

(b) Explain why $\hat{\gamma}_1 \xrightarrow{p} 0$ when the initial model $\hat{\mu}_1^{(0)}$ is correctly specified, and what this implies about the efficiency cost of including the clever covariate.

(c) Explain in words why $\hat{\pi}_i^{-1}$ is called the clever covariate: what bias does it absorb, and why does this make the prediction estimator a debiased estimator even when $\hat{\mu}_1^{(0)}$ is misspecified?

7. Coverage of the Wald confidence interval (computational). Use the DGP of Section 12.10 with the propensity score generalized to $\pi^*(x) = \text{expit}\{0.2x + \gamma(x^2 - 1)\}$, so that the baseline ($\gamma = 0.2$) coincides with the lab and larger γ degrades overlap. Implement the AIPW estimator and variance estimator Equation 11.17. Following Scenario S3, use logistic regression of T on $(1, X, X^2)$ for $\hat{\pi}$ (correct PS) and OLS of Y on $(1, X)$ separately in each arm for $\hat{\mu}_t$ (misspecified OR); truncate $\hat{\pi}$ to $[10^{-3}, 1-10^{-3}]$.

(a) With $\gamma = 0.2$, $n = 500$, $B = 2000$ replications, compute the empirical coverage of the 95% Wald interval $\hat{\tau}_{\text{AIPW}} \pm z_{0.975}\sqrt{\widehat{V}}$. Compare the average SE $\overline{\sqrt{\widehat{V}}}$ with the empirical SD of $\hat{\tau}_{\text{AIPW}}$ across replications.

(b) Repeat with $\gamma \in \{0.5, 1.0\}$, producing increasingly weak overlap. Report coverage, average SE, and empirical SD. What do you observe?

(c) Which assumption of Section 13.6 is stressed as γ grows, and which of the remedies in the Extreme Propensity Scores warning would you try first? (No simulation required; a paragraph suffices.)

Chapter 12

Flexible Nuisance Estimation, Orthogonal Scores, and Cross-Fitting

Learning Objectives

By the end of this chapter, students should be able to:

1. Explain why flexible nuisance estimation can produce misleading inference for a causal parameter even when predictive accuracy is high.
2. Define Neyman orthogonality, verify it for a given estimating function, and explain how it reduces sensitivity to first-order nuisance estimation error.
3. Identify the orthogonal score for the ATE, recognize it as the efficient influence function from Chapter 11, and state the three roles it simultaneously fulfills.
4. Describe sample splitting, explain why independence between nuisance estimation and score evaluation simplifies asymptotic arguments, and articulate its efficiency cost.
5. Describe K-fold cross-fitting, write the cross-fitted moment equation explicitly, and explain how cross-fitting recovers efficiency relative to a single split.
6. State the product-rate condition for the cross-fitted AIPW estimator, connect it to Neyman orthogonality, and explain why it allows each nuisance estimator to converge at rate $n^{-1/4}$.
7. Construct a consistent variance estimator from out-of-fold estimated influence values and form a Wald confidence interval.
8. Articulate what machine learning can and cannot contribute to causal inference, and follow the practical workflow for cross-fitted orthogonal-score estimation.

12.1 Why Flexible Nuisance Estimation Is Both Attractive and Dangerous

In Chapters 10 and 11 we developed estimation and inference using estimating equations, influence functions, doubly robust scores, and semiparametric efficiency. A central feature of these methods is that the causal parameter depends on nuisance functions such as the outcome regressions $\mu_t(x) = \mathbb{E}(Y \mid T{=}t, X{=}x)$ and the propensity score $\pi(x) = P(T{=}1 \mid X{=}x)$. When these nuisance functions are too complex for low-dimensional parametric forms, flexible methods — penalized regression, splines, random forests, boosting, neural networks, ensembles — become attractive.

These methods can substantially improve predictive accuracy, but they create new statistical difficulties: slower convergence, overfitting, difficult asymptotic characterization, and failure of naive plug-in inference even when prediction quality is high.

Better nuisance prediction therefore does not automatically imply valid inference for the target causal parameter. This chapter explains how orthogonal scores, sample splitting, and cross-fitting make it possible to combine flexible nuisance estimation with valid large-sample inference. The solution combines an *orthogonal score*, which removes first-order sensitivity to nuisance estimation error, with *cross-fitting*,

which separates nuisance estimation from score evaluation to control the remaining empirical process remainder.

12.2 Why Naive Plug-In Estimation Can Fail

Suppose $\psi = \Psi(\eta)$, where η denotes a nuisance object such as (μ_0, μ_1, π). The *plug-in estimator* is $\hat{\psi}_{\text{plug}} = \Psi(\hat{\eta})$.

12.2.1 The First-Order Taylor Expansion

Apply a first-order Taylor expansion of Ψ around the true nuisance η_0:

$$\hat{\psi}_{\text{plug}} - \psi = \underbrace{D_\eta \Psi(\eta_0)[\hat{\eta} - \eta_0]}_{\text{linear term}} + \underbrace{R(\hat{\eta}, \eta_0)}_{\text{second-order remainder}}, \tag{12.1}$$

where $D_\eta \Psi(\eta_0)[h]$ denotes the Gateaux derivative and $|R(\hat{\eta}, \eta_0)| \lesssim \|\hat{\eta} - \eta_0\|^2$.

The second-order remainder is manageable: if $\|\hat{\eta} - \eta_0\| = o_p(n^{-1/4})$ then $R = o_p(n^{-1/2})$. The obstacle is the *linear term*, proportional to the nuisance estimation error $\hat{\eta} - \eta_0$.

12.2.2 Finite-Dimensional Nuisance: Orthogonality Is Sufficient

When $\eta \in \mathbb{R}^d$ and $\hat{\eta} - \eta_0 = O_p(n^{-1/2})$, the linear term in Equation 12.1 is $O_p(n^{-1/2})$ and contributes to the limiting distribution. If the functional satisfies $D_\eta \Psi(\eta_0) = 0$ — the *Neyman orthogonality* condition of Section 12.3 — then the linear term vanishes. The expansion reduces to the second-order remainder alone, which at parametric rates is $O_p(n^{-1})$ and hence negligible. In the finite-dimensional setting, **Neyman orthogonality is sufficient** to remove nuisance contamination.

12.2.3 Infinite-Dimensional Nuisance: Orthogonality Is Not Enough

When η belongs to a function space, orthogonality is no longer sufficient on its own. Write the plug-in estimating equation as $\mathbb{P}_n\{\phi(\,\cdot\,; \psi, \hat{\eta})\} = 0$, and decompose the deviation from the population equation:

$$\mathbb{P}_n\{\phi(\,\cdot\,; \psi, \hat{\eta})\} - \mathbb{P}_n\{\phi(\,\cdot\,; \psi, \eta_0)\} = \underbrace{P\{\phi(\,\cdot\,; \psi, \hat{\eta}) - \phi(\,\cdot\,; \psi, \eta_0)\}}_{\text{population bias}} + \underbrace{(\mathbb{P}_n - P)\{\phi(\,\cdot\,; \psi, \hat{\eta}) - \phi(\,\cdot\,; \psi, \eta_0)\}}_{\text{empirical process term}}. \tag{12.2}$$

Neyman orthogonality controls the *population bias*: the Gateaux derivative of $P\phi$ vanishes at the truth, so the population bias is second order. It says nothing about the *empirical process term*, whose behavior depends on the complexity of the function class $\{\phi(\,\cdot\,; \psi, \eta) : \eta \in \mathcal{H}\}$. For flexible machine-learning estimators this class is typically too rich for classical empirical-process arguments, and the term can remain non-negligible even as $\hat{\eta}$ converges consistently.

> Remark
>
> This problem is not specific to machine learning. It arises whenever nuisance parameters are estimated in an infinite-dimensional model. Machine learning makes the issue more visible because it encourages highly adaptive nuisance estimation; see Chernozhukov et al. (2018) for a systematic treatment.

12.3 Orthogonal Scores

> Remark: From Functionals to Moment Equations
>
> Section 12.2 formulated orthogonality at the level of a functional as $D_\eta \Psi(\eta_0) = 0$. Most practical estimators are solutions to a moment equation $\mathbb{P}_n\{\phi(\,\cdot\,; \hat{\psi}, \hat{\eta})\} = 0$, so it is more useful to phrase orthogonality in terms of the score ϕ. By the implicit-function theorem, the parameter $\psi(\eta) = \operatorname{argzero}_\psi P\{\phi(\,\cdot\,; \psi, \eta)\}$ satisfies $D_\eta \psi(\eta_0)[h] = -[\partial_\psi P\phi]^{-1}\, \partial_r P\{\phi(\,\cdot\,; \psi_0, \eta_0 + rh)\}|_{r=0}$, so vanishing

of the nuisance Gateaux derivative of $P\phi$ is equivalent to vanishing of $D_\eta\psi$. The two formulations express the same orthogonality condition in two languages.

Definition: Neyman Orthogonality

The score function $\phi(O;\, \psi, \eta)$ is called **orthogonal** (or **Neyman orthogonal**) at the truth (ψ_0, η_0) if the Gateaux derivative of the estimating equation with respect to the nuisance, evaluated at the truth, vanishes:

$$\frac{\partial}{\partial r}\mathbb{E}\{\phi(O;\, \psi_0,\, \eta_0 + rh)\}\Big|_{r=0} = 0 \tag{12.3}$$

for all perturbation directions h in a suitable class.

This condition means that small local perturbations of the nuisance have no first-order effect on the estimating equation at the truth. Orthogonality controls only the population-level bias; a separate argument is needed for the empirical process term when the same sample is reused.

Remark: Neyman's Original Insight

The terminology honors Jerzy Neyman, whose work on hypothesis testing introduced the idea of constructing test statistics insensitive to nuisance parameters. In the modern semiparametric literature the concept was formalized by Robins et al. (1994) and others, and later made systematic in the double machine learning framework of Chernozhukov et al. (2018).

12.4 The Orthogonal Score for the Average Treatment Effect

Under consistency, conditional exchangeability, and positivity, the efficient influence function from Chapter 11 is:

$$\varphi_{\text{eff}}(O;\, \tau, \eta) = \frac{T}{\pi(X)}\{Y - \mu_1(X)\} - \frac{1-T}{1-\pi(X)}\{Y - \mu_0(X)\} + \mu_1(X) - \mu_0(X) - \tau, \tag{12.4}$$

where $\eta = (\mu_0, \mu_1, \pi)$. Chapter 11 identified this as the right score for the ATE; Chapter 12's role is to explain how to estimate its nuisance components safely with flexible learners.

Lemma: Neyman Orthogonality of the ATE Score

The score $\varphi_{\text{eff}}(O;\, \tau, \eta)$ in Equation 12.4 is Neyman orthogonal at every truth (τ_0, η_0) satisfying the identification assumptions. That is, for every bounded perturbation direction $h(x)$, $g(x)$, or $\ell(x)$:

$$\frac{\partial}{\partial r}\mathbb{E}\{\varphi_{\text{eff}}(O;\, \tau_0,\, \mu_0,\, \mu_1,\, \pi + rh)\}\Big|_{r=0} = 0,$$

and likewise for perturbations in μ_1 and μ_0.

Proof

Differentiation and expectation are interchanged under standard bounded-convergence arguments; the LIE is applied by conditioning on X. The defining identity $\mathbb{E}(T \mid X) = \pi(X)$ is the engine of all three calculations: it implies the two **balancing identities**

$$\mathbb{E}\left(\frac{T}{\pi(X)}\,\middle|\, X\right) = 1, \qquad \mathbb{E}\left(\frac{1-T}{1-\pi(X)}\,\middle|\, X\right) = 1, \tag{12.5}$$

and consequently $\mathbb{E}\{T(Y-\mu_1(X)) \mid X\} = 0$ and $\mathbb{E}\{(1-T)(Y-\mu_0(X)) \mid X\} = 0$. Each perturbation derivative reduces to one of these conditional expectations.

Perturbation in π. Replace π by $\pi + rh$ and differentiate. The only π-dependent terms are the two IPW residual terms:

$$\frac{\partial}{\partial r}\Big|_{r=0} = \mathbb{E}\left[-\frac{T\,h(X)}{\pi(X)^2}\{Y-\mu_1(X)\} - \frac{(1-T)\,h(X)}{(1-\pi(X))^2}\{Y-\mu_0(X)\}\right] = 0$$

by the two residual identities above (factor out $h(X)/\pi(X)^2$ and $h(X)/(1-\pi(X))^2$, then condition on X).
Perturbation in μ_1. Replace μ_1 by $\mu_1 + rg$. Only $-(T/\pi(X))\mu_1(X)$ and $\mu_1(X)$ depend on μ_1, so:

$$\mathbb{E}\left[g(X)\left(1-\frac{T}{\pi(X)}\right)\right] = 0$$

by the first identity in Equation 12.5.
Perturbation in μ_0. Replace μ_0 by $\mu_0 + r\ell$. Symmetrically:

$$\mathbb{E}\left[\ell(X)\left(\frac{1-T}{1-\pi(X)}-1\right)\right] = 0$$

by the second identity in Equation 12.5. □

Remark: What the Proof Reveals

The two balancing identities Equation 12.5 are immediate consequences of the defining identity $\mathbb{E}(T \mid X) = \pi(X)$, which is also the engine behind the balancing theorem and the IPW identification formula (Chapter 6). Neyman orthogonality of the ATE score is therefore a direct consequence of how the propensity score is defined, not an additional requirement imposed on it.

The score Equation 12.4 simultaneously fulfills three purposes:

1. It *identifies* τ through the moment condition $\mathbb{E}\{\varphi_{\text{eff}}(O;\,\tau,\eta)\} = 0$.
2. It is *doubly robust*, yielding a consistent estimator whenever either the outcome model or the propensity score model is correctly specified.
3. It is the *efficient influence function*, so an estimator that is regular, asymptotically linear with this influence function, and whose nuisance remainders are asymptotically negligible achieves the semiparametric efficiency bound.

Remark: Double Robustness versus Semiparametric Efficiency

Properties (ii) and (iii) are distinct claims requiring distinct conditions. *Double robustness* (ii) is a consistency claim: under standard overlap and LLN regularity, the AIPW estimator is consistent for τ if either $\hat{\mu}_t$ converges to the true μ_t *or* $\hat{\pi}$ converges to the true π, but not necessarily both. *Semiparametric efficiency* (iii) is a distributional claim about $\sqrt{n}(\hat{\tau}-\tau_0)$: it requires that both nuisance estimators converge to the truth at rates satisfying the product-rate condition Equation 12.12. When only one nuisance block is correctly specified, the estimator can remain consistent by double robustness, but the asymptotic-linearity expansion with influence function φ_{eff} and the efficient-bound variance need not apply, and inference based on them can be misleading without further analysis of the actual limiting estimating function.

12.5 Why Reusing the Same Data Can Be Problematic

12.5.1 The Plug-In Remainder for the AIPW Estimator

Consider the plug-in AIPW estimator defined by the moment equation $\mathbb{P}_n\{\varphi_{\text{eff}}(O;\,\hat{\tau}_{\text{plug}},\hat{\eta})\} = 0$ where the same sample is used for both nuisance estimation and score evaluation. Because φ_{eff} is linear in τ with $\partial_\tau \varphi_{\text{eff}} = -1$:

$$\hat{\tau}_{\text{plug}} - \tau_0 = \mathbb{P}_n\{\varphi_{\text{eff}}(O;\,\tau_0,\eta_0)\} + \underbrace{\mathbb{P}_n\{\varphi_{\text{eff}}(O;\,\tau_0,\hat{\eta}) - \varphi_{\text{eff}}(O;\,\tau_0,\eta_0)\}}_{=:R_n}. \tag{12.6}$$

The first term satisfies the CLT. The remainder R_n decomposes as in Equation 12.2:

$$R_n = \underbrace{P\{\varphi_{\text{eff}}(\,\cdot\,;\tau_0,\hat{\eta}) - \varphi_{\text{eff}}(\,\cdot\,;\tau_0,\eta_0)\}}_{\text{population bias}} + \underbrace{(\mathbb{P}_n - P)\{\varphi_{\text{eff}}(\,\cdot\,;\tau_0,\hat{\eta}) - \varphi_{\text{eff}}(\,\cdot\,;\tau_0,\eta_0)\}}_{\text{empirical process term}}. \tag{12.7}$$

By the Neyman orthogonality lemma, the population bias is $o_p(n^{-1/2})$ under the product-rate condition. Writing $f_\eta(\cdot) = \varphi_{\text{eff}}(\,\cdot\,;\tau_0,\eta)$, the empirical process term takes the form:

$$(\mathbb{P}_n - P)\{f_{\hat{\eta}} - f_{\eta_0}\} = \frac{1}{\sqrt{n}}\,\mathbb{G}_n\{f_{\hat{\eta}} - f_{\eta_0}\}. \tag{12.8}$$

This term depends on $\hat{\eta}$, which is estimated from the same sample. Whether it is $o_p(n^{-1/2})$ depends on how complex the class $\{f_\eta : \eta \in \mathcal{H}\}$ is — a question the Donsker condition answers.

12.5.2 The Donsker Condition

Definition: Donsker Class

A class of measurable functions $\mathcal{F}$ is a **Donsker class** (with respect to P) if the empirical process $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges weakly in $\ell^\infty(\mathcal{F})$ to a tight Gaussian process. In particular, every Donsker class satisfies:

$$\sup_{f\in\mathcal{F}} \left|(\mathbb{P}_n - P)f\right| = O_p(n^{-1/2}),$$

and the empirical process $\mathbb{G}_n$ is stochastically equicontinuous over $\mathcal{F}$. See Vaart (1998), Chapter 19, for a complete treatment.

A key sufficient condition is finite bracketing entropy: the class $\mathcal{F}$ is Donsker whenever $\int_0^{\delta_0} \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))}\, d\epsilon < \infty$, where $N_{[]}$ counts the minimum number of ϵ-brackets needed to cover $\mathcal{F}$.

Example: Donsker and Non-Donsker Classes

- **Parametric classes** (e.g., logistic regression indexed by a finite-dimensional parameter) are Donsker under mild moment conditions.
- **Hölder-smooth function classes** on $[0,1]^d$ with smoothness index $s > d/2$ are Donsker.
- **Indicator classes** $\{\mathbf{1}(x \le t) : t \in \mathbb{R}\}$ are Donsker by the classical Donsker theorem.
- **Random forests and neural networks** in common modern implementations typically fall outside classical fixed Donsker regimes, especially when complexity grows with n.
- **High-dimensional lasso** with a growing number of selected variables falls outside fixed Donsker regimes; sparsity-based empirical-process arguments are typically used instead.

12.5.3 Connecting the Two Terms

Returning to Equation 12.7:

- **Population bias.** Controlled by Neyman orthogonality alone; $o_p(n^{-1/2})$ under the product-rate condition regardless of how $\hat{\eta}$ is estimated.
- **Empirical process term.** If $\{f_\eta : \eta \in \mathcal{H}\}$ is Donsker and $\|\hat{\eta} - \eta_0\| \to 0$ in probability, then by stochastic equicontinuity, $\mathbb{G}_n\{f_{\hat{\eta}} - f_{\eta_0}\} = o_p(1)$, making the term $o_p(n^{-1/2})$.

When the Donsker condition fails, the empirical process term can remain non-negligible even as $\hat{\eta} \to \eta_0$. **Orthogonality and cross-fitting are complements, not substitutes**: orthogonality controls the population bias; cross-fitting controls the empirical process term.

Remark: Donsker Conditions and Machine Learning

Common modern implementations of random forests, neural networks, and high-dimensional regularized learners often fall outside classical fixed Donsker regimes. Cross-fitting addresses this by changing the dependence structure: because $\hat{\eta}^{(-k)}$ is independent of the observations in $\mathcal{I}_k$, the

empirical process term can be controlled by a conditional argument instead of a Donsker assumption.

12.6 Sample Splitting

Sample splitting resolves the empirical process problem by construction. Partition $\{1, \dots, n\}$ into a training sample $\mathcal{I}_{\text{train}}$ and an evaluation sample $\mathcal{I}_{\text{eval}}$. Fit nuisance estimators $\hat{\eta}^{(\text{train})}$ on $\mathcal{I}_{\text{train}}$, then solve the estimating equation on the held-out sample:

$$\frac{1}{|\mathcal{I}_{\text{eval}}|} \sum_{i \in \mathcal{I}_{\text{eval}}} \varphi_{\text{eff}}(O_i;\, \tau,\, \hat{\eta}^{(\text{train})}) = 0.$$

The remainder on the evaluation fold decomposes as:

$$R_n^{(\text{e})} = \underbrace{P\{\varphi_{\text{eff}}(\,\cdot\,;\tau, \hat{\eta}^{(\text{train})}) - \varphi_{\text{eff}}(\,\cdot\,;\tau, \eta_0)\}}_{o_p(n^{-1/2}) \text{ by orthogonality}} + \underbrace{(\mathbb{P}_{n_e} - P)\{\varphi_{\text{eff}}(\,\cdot\,;\tau, \hat{\eta}^{(\text{train})}) - \varphi_{\text{eff}}(\,\cdot\,;\tau, \eta_0)\}}_{\text{empirical process term}}. \tag{12.9}$$

Because $\hat{\eta}^{(\text{train})}$ is independent of the observations in $\mathcal{I}_{\text{eval}}$, conditioning on $\hat{\eta}^{(\text{train})}$ makes the summands conditionally i.i.d. with mean zero. Under L_2 consistency, strong overlap, and finite-variance moment conditions, a conditional variance bound gives the empirical process term $= o_p(n^{-1/2})$. **No Donsker condition is needed**: independence allows a conditional argument in place of stochastic equicontinuity.

The main disadvantage is inefficiency: only $|\mathcal{I}_{\text{eval}}|$ observations contribute to the estimation of τ, and the estimate depends on the particular random partition chosen.

12.7 Cross-Fitting

Cross-fitting extends sample splitting to recover full-sample efficiency while preserving the independence argument. Partition $\{1, \dots, n\}$ into K approximately equal folds $\mathcal{I}_1, \dots, \mathcal{I}_K$. For each fold k: estimate nuisance functions on the complement $\hat{\eta}^{(-k)} = (\hat{\mu}_0^{(-k)}, \hat{\mu}_1^{(-k)}, \hat{\pi}^{(-k)})$, then evaluate the score on the held-out fold. The cross-fitted estimator $\hat{\tau}_{\text{DML}}$ solves:

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \varphi_{\text{eff}}(O_i;\, \tau,\, \hat{\eta}^{(-k)}) = 0. \tag{12.10}$$

For each fold k, the nuisance estimate $\hat{\eta}^{(-k)}$ is trained on $\{1, \dots, n\} \setminus \mathcal{I}_k$, which is independent of $\mathcal{I}_k$. Conditioning on $\hat{\eta}^{(-k)}$, the fold-k summands are independent and mean zero. The fold-by-fold argument applies to each $R_n^{(k)}$ separately, giving $\sqrt{n} R_n^{(k)} = o_p(1)$ for each k without any Donsker assumption.

Remark: Independence Structure of Cross-Fitting

The bound on each $R_n^{(k)}$ uses only that $\hat{\eta}^{(-k)}$ is independent of its own evaluation fold $\mathcal{I}_k$, which holds by construction. The nuisance estimates $\hat{\eta}^{(-1)}, \dots, \hat{\eta}^{(-K)}$ are *not* mutually independent — any two share most of their training observations — but cross-fold independence is never invoked. Because the $o_p(1)$ bound on each $R_n^{(k)}$ is uniform across k for fixed K, averaging over folds preserves the bound.

Cross-fitting achieves two goals simultaneously: it preserves the held-out independence structure that removes the need for Donsker conditions, and ensures every observation serves once as an evaluation point so the full sample contributes to the estimation of τ.

Algorithm: Cross-Fitted Orthogonal-Score Estimator for the ATE

1. Partition the sample into K folds $\mathcal{I}_1, \dots, \mathcal{I}_K$.
2. For each $k = 1, \dots, K$, fit nuisance estimators $\hat{\eta}^{(-k)} = (\hat{\mu}_0^{(-k)}, \hat{\mu}_1^{(-k)}, \hat{\pi}^{(-k)})$ on the complement

$\{1, \dots, n\} \setminus \mathcal{J}_k$.

3. For each observation $i \in \mathcal{J}_k$, compute the out-of-fold score contribution $\hat{\varphi}_i = \varphi_{\text{eff}}(O_i; \tau, \hat{\eta}^{(-k)})$.
4. Solve the cross-fitted moment equation Equation 12.10. Because the score is linear in τ, the solution is:

$$\hat{\tau}_{\text{DML}} = \hat{\mu}_{1,\text{DML}} - \hat{\mu}_{0,\text{DML}}, \tag{12.11}$$

where $\hat{\mu}_{t,\text{DML}} = n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{J}_k} [\hat{\mu}_t^{(-k)}(X_i) + (t \cdot T_i + (1-t)(1-T_i))/\hat{\pi}_t^{(-k)}(X_i) \cdot \{Y_i - \hat{\mu}_t^{(-k)}(X_i)\}]$.

12.8 Double Machine Learning for the Average Treatment Effect

The estimator Equation 12.11 is simply the AIPW estimator from Chapter 11, with out-of-fold nuisance estimates substituted in place of full-sample estimates. The term *double machine learning* (DML), introduced by Chernozhukov et al. (2018), refers to the combination of three ingredients: (i) *Neyman orthogonality*, which reduces the population bias to a second-order product; (ii) *cross-fitting*, which removes same-sample dependence in the empirical-process term; and (iii) *flexible machine-learning estimation* of the nuisance functions, made valid by the first two ingredients.

The word "double" should not be read as merely meaning "two models"; it refers to the combination of nuisance learning with orthogonalization and debiasing. DML is not a new causal estimand and not a new identification strategy; it is a modern estimation protocol built around an orthogonal score. The cross-fitted AIPW estimator in the biostatistics literature is the same object.

"Double" Does Not Mean "Two Models Are Enough"

It is tempting to think that estimating any two of μ_0, μ_1, and π flexibly is sufficient. What matters is that *all* nuisance components entering the orthogonal score are estimated out-of-fold, and that the product-rate condition is satisfied. Using flexible methods for only one nuisance function while misspecifying another can still produce invalid inference.

12.9 Rate Conditions and Asymptotic Normality

Theorem: Asymptotic Linearity Under Cross-Fitting

Suppose: (i) the score $\varphi_{\text{eff}}(O; \tau, \eta)$ is orthogonal at the truth; (ii) the nuisance estimators are consistent in mean squared error; (iii) **strong overlap**: there exists $c > 0$ such that $c \le \pi(X) \le 1-c$ and $c \le \hat{\pi}^{(-k)}(X) \le 1 - c$ a.s. for each fold k; (iv) the **product-rate condition**:

$$\|\hat{\pi}^{(-k)} - \pi\| \cdot \|\hat{\mu}_t^{(-k)} - \mu_t\| = o_p(n^{-1/2}), \qquad t = 0, 1. \tag{12.12}$$

Then the cross-fitted estimator satisfies:

$$\sqrt{n}(\hat{\tau}_{\text{DML}} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{\text{eff}}(O_i; \tau, \eta) + o_p(1) \xrightarrow{d} N\Big(0,\ \mathbb{E}[\varphi_{\text{eff}}(O; \tau, \eta)^2]\Big).$$

The asymptotic variance equals the semiparametric efficiency bound; $\hat{\tau}_{\text{DML}}$ is asymptotically efficient whenever Equation 12.12 holds.

Remark: Weak Overlap vs. Strong Overlap

Weak overlap ($0 < \pi(X) < 1$ a.s.) is the condition needed for *identification*. *Strong overlap* ($c \le \pi(X) \le 1 - c$) is what the proof actually uses: it bounds the inverse weights away from infinity, ensuring the Cauchy-Schwarz step in Step 1 and the variance bound in Step 2 go through. Without strong overlap, influence function values can have infinite variance and $\sqrt{n}$-inference breaks down even if the ATE is identified (Khan and Tamer 2010). The requirement on $\hat{\pi}^{(-k)}$ is equally important: even if the true π is well-behaved, an estimated propensity score that strays near 0 or 1 in a given

fold will cause the same instability.

Proof Sketch

Write $n_k = |\mathcal{I}_k|$ and $\mathbb{P}_{n_k}$ for the empirical measure over fold k. From the cross-fitted moment equation Equation 12.10:

$$\hat{\tau}_{\mathrm{DML}} - \tau = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathrm{eff}}(O_i;\, \tau, \eta_0) + \frac{1}{K}\sum_{k=1}^{K} R_n^{(k)},$$

where each fold remainder $R_n^{(k)}$ decomposes as in Equation 12.9. It suffices to show $\sqrt{n}\, R_n^{(k)} = o_p(1)$ for each k.

Step 1: Population bias. By iterated expectations conditioning on X:

$$B_n^{(k)} = -\mathbb{E}\left[(\hat{\mu}_1^{(-k)} - \mu_1)(X) \cdot \frac{\hat{\pi}^{(-k)}(X) - \pi(X)}{\hat{\pi}^{(-k)}(X)} - (\hat{\mu}_0^{(-k)} - \mu_0)(X) \cdot \frac{\hat{\pi}^{(-k)}(X) - \pi(X)}{1 - \hat{\pi}^{(-k)}(X)}\right].$$

Under strong overlap, the denominators are bounded below. By Cauchy-Schwarz:

$$|B_n^{(k)}| \le C\Big(\|\hat{\mu}_1^{(-k)} - \mu_1\| \cdot \|\hat{\pi}^{(-k)} - \pi\| + \|\hat{\mu}_0^{(-k)} - \mu_0\| \cdot \|\hat{\pi}^{(-k)} - \pi\|\Big) = o_p(n^{-1/2})$$

by the product-rate condition Equation 12.12. Hence $\sqrt{n}\, B_n^{(k)} = o_p(1)$.

Step 2: Empirical process term. Condition on $\hat{\eta}^{(-k)}$. Since $\hat{\eta}^{(-k)}$ is trained on $\{1, \dots, n\} \setminus \mathcal{I}_k$, the summands for $i \in \mathcal{I}_k$ are conditionally i.i.d. with conditional mean zero (after removing $B_n^{(k)}$). Let $f_k = \varphi_{\mathrm{eff}}(\cdot; \tau, \hat{\eta}^{(-k)}) - \varphi_{\mathrm{eff}}(\cdot; \tau, \eta_0)$. The conditional variance:

$$\mathrm{Var}(E_n^{(k)} \mid \hat{\eta}^{(-k)}) = \frac{1}{n_k}\, \|f_k\|_{L_2}^2.$$

Under strong overlap and consistency, $\|f_k\|_{L_2}^2 = o_p(1)$, so $\mathrm{Var}(\sqrt{n}\, E_n^{(k)} \mid \hat{\eta}^{(-k)}) = (n/n_k)\, \|f_k\|_{L_2}^2 = O(1) \cdot o_p(1) = o_p(1)$. By conditional Chebyshev, $\sqrt{n}\, E_n^{(k)} = o_p(1)$.

Conclusion. Combining Steps 1 and 2, $\sqrt{n}\, R_n^{(k)} = o_p(1)$ for each k, and hence $\sqrt{n} \cdot K^{-1} \sum_k R_n^{(k)} = o_p(1)$. The first term converges in distribution to $N(0, \mathbb{E}[\varphi_{\mathrm{eff}}^2])$ by the ordinary CLT. Asymptotic linearity and normality follow by Slutsky. □

Remark: Interpreting the Product-Rate Condition

Condition Equation 12.12 requires the *product* of the two nuisance estimation errors to be $o_p(n^{-1/2})$. A sufficient symmetric condition is that each nuisance estimator converges at rate $o_p(n^{-1/4})$ in $L_2(P)$ norm. This rate is achievable by many nonparametric and machine-learning methods under regularity conditions, and is what makes flexible nuisance estimation feasible in semiparametric causal inference.

Remark: Connection to Chapter 11

?@thm-asymp is the cross-fitting version of the asymptotic normality result in Chapter 11. The product-rate condition Equation 12.12 is identical to the one in Chapter 11; the difference is that here the nuisance estimates are out-of-fold, which removes the need for Donsker conditions. Cross-fitting provides the same asymptotic conclusion while accommodating a substantially broader class of nuisance learners.

12.10 Lab: Simulation Study of the DML Estimator

This lab verifies the theoretical properties of **?@thm-asymp**. The central message is that flexible nuisance estimation alone is not sufficient for valid inference: a consistent nonparametric estimator can still produce a biased AIPW estimator if the same sample is reused for nuisance estimation and score evaluation.

DGP. $n = 2000$, covariates $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \text{Uniform}(-1, 1)$. True nuisance functions:

$$\pi(X) = \text{expit}(\sin(\pi X_1) + X_2^2 - 0.5), \quad \mu_1(X) = 2\sin(\pi X_1) + X_2^2 + X_3, \quad \mu_0(X) = \sin(\pi X_1) + X_2^2 - X_3.$$

True ATE: $\tau = \mathbb{E}\{\mu_1(X) - \mu_0(X)\} = \mathbb{E}\{\sin(\pi X_1) + 2X_3\} = 0$ (both X_1 and X_3 symmetric about zero). The nonlinearity of π and μ_t means random forests can fit them consistently; their function class falls outside classical fixed Donsker regimes.

Three estimators. (1) *Oracle AIPW*: plug in the true nuisance functions — infeasible but provides the semiparametric efficiency benchmark. (2) *Naive RF AIPW*: estimate nuisance functions using random forests on the *full sample*, then evaluate the AIPW score on the same sample. (3) *DML (cross-fitted AIPW)*: same random forest learners, but via $K = 5$ fold cross-fitting. Estimators (2) and (3) use identical learners; the only difference is data reuse versus cross-fitting.

Settings. Software: Python, scikit-learn. Replications: $n_{\text{sim}} = 200$. Folds: $K = 5$, shuffled. Nuisance learners: random forest regressor for μ_t; random forest classifier for π; `n_estimators=100`, `min_samples_leaf=5`. Propensity trimming: $\hat{\pi}$ clipped to $[0.01, 0.99]$. No hyperparameter tuning inside the cross-fitting loop.

Results:

Estimator	Bias	Variance	RMSE
Oracle AIPW	0.009	0.0029	0.055
Naive RF AIPW	0.039	0.0031	0.068
DML	0.007	0.0036	0.060

Oracle AIPW is nearly unbiased with small variance — the semiparametric efficiency benchmark.

Naive RF AIPW exhibits a clear positive bias (0.039, roughly four times the oracle bias) despite using consistent random forest estimators. The bias arises from the uncontrolled empirical process term Equation 12.8: reusing the same sample allows overfitting in the random forests to propagate into the AIPW score.

DML reduces the bias to 0.007 — within Monte Carlo error of zero and indistinguishable from the oracle — at a modest finite-sample variance cost (from 0.0031 to 0.0036). The small variance increase is a finite-sample phenomenon: each out-of-fold nuisance estimator is trained on $(K-1)n/K = 1600$ observations, so its predictions are slightly noisier than in-sample fits. The Naive estimator's lower variance is partly a consequence of overfitting. The net effect is a lower RMSE for DML (0.060) than for Naive RF AIPW (0.068): the bias reduction outweighs the variance difference.

> Remark: The Key Comparison
>
> The key comparison is between Naive RF AIPW and DML, not between either and the oracle. Both use identical random forest learners; the only difference is cross-fitting. The bias reduction from 0.039 to 0.007 is attributable solely to removing the same-sample data-reuse problem. The accompanying small increase in variance is a finite-sample phenomenon, not an asymptotic penalty: the gap should narrow as n grows.

12.11 Variance Estimation and Confidence Intervals

Inference for $\hat{\tau}_{\text{DML}}$ proceeds exactly as in Chapter 11, except that nuisance estimates are now out-of-fold. Let $k(i)$ denote the fold containing observation i, and define the estimated influence value:

$$\hat{\varphi}_i = \varphi_{\text{eff}}(O_i; \hat{\tau}_{\text{DML}}, \hat{\eta}^{(-k(i))}).$$

The variance estimator is:

$$\hat{V} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\varphi}_i - \bar{\varphi})^2, \qquad \bar{\varphi} = \frac{1}{n} \sum_{i=1}^{n} \hat{\varphi}_i. \tag{12.13}$$

The Wald confidence interval is $\hat{\tau}_{\text{DML}} \pm z_{1-\alpha/2} \sqrt{\hat{V}}$.

Theorem: Consistency of the Variance Estimator

Under the conditions of **?@thm-asymp**, the variance estimator $\hat{V}$ in Equation 12.13 is consistent for the asymptotic variance: $n\hat{V} \xrightarrow{p} \mathbb{E}[\varphi_{\text{eff}}(O; \tau, \eta)^2]$. The Wald confidence interval is therefore asymptotically valid.

The formula is identical to the one in Chapter 11; the only difference is that $\hat{\varphi}_i$ uses the out-of-fold nuisance estimate $\hat{\eta}^{(-k(i))}$ rather than a full-sample estimate.

Remark: Centering Is Numerically Inconsequential

Because φ_{eff} is linear in τ with $\partial_\tau \varphi_{\text{eff}} = -1$, evaluating the score at the solution $\hat{\tau}_{\text{DML}}$ of Equation 12.10 forces $\bar{\varphi} = n^{-1} \sum_{i=1}^{n} \hat{\varphi}_i = 0$ exactly (Exercise 5). The centering is included only because the formula then matches the conventional sample-variance expression and remains stable under finite-precision arithmetic.

12.12 A Practical Workflow

Workflow: Cross-Fitted Causal Estimation

Phase 1: Identification and Setup

1. **Specify the estimand.** State the target causal parameter and verify the identification assumptions — conditional exchangeability, positivity, and consistency for the ATE.
2. **Choose the orthogonal score.** For the ATE, this is the efficient influence function Equation 12.4. The score fixes which nuisance components must be estimated.
3. **Select nuisance learners.** Choose flexible learners for $\mu_0(x)$, $\mu_1(x)$, and $\pi(x)$. Verify informally that the product-rate condition Equation 12.12 is plausible.
4. **Assess covariate overlap.** Before fitting, examine the empirical distribution of each covariate by treatment group. Near-violations of positivity at the design level will inflate influence values and destabilize inference.

Phase 2: Estimation

5. **Partition and cross-fit.** Partition into K folds ($K = 5$ is a common default). For each fold k, fit nuisance learners on the complement.
6. **Solve the moment equation.** Apply the explicit formula Equation 12.11 with out-of-fold nuisance estimates.
7. **Compute estimated influence values.** For each $i \in \mathcal{I}_k$, compute $\hat{\varphi}_i = \varphi_{\text{eff}}(O_i; \hat{\tau}_{\text{DML}}, \hat{\eta}^{(-k)})$.
8. **Estimate the variance.** Compute $\hat{V}$ from Equation 12.13 and form the Wald confidence interval.

Phase 3: Diagnostics and Interpretation

9. **Inspect estimated propensity scores.** Examine the distribution of $\hat{\pi}^{(-k)}(X_i)$ across folds; extreme values near 0 or 1 inflate influence-value variance.
10. **Assess sensitivity.** Repeat the analysis with alternative learners. Substantial sensitivity to learner choice signals the product-rate condition may not be satisfied at this sample size.
11. **Interpret relative to identification assumptions.** The credibility of the causal estimate rests on the identification assumptions, not on predictive accuracy. These assumptions are untestable from the data.

Step 11 is the most important step in the workflow. Identification sits outside what any algorithm can verify.

12.13 What Machine Learning Does Not Solve

Flexible nuisance estimation is a genuine improvement over parametric methods, but it does not resolve the fundamental difficulties of causal inference. Machine learning does not solve:

- **Unmeasured confounding.** If important confounders are absent from X, conditional exchangeability $Y(t) \perp\!\!\!\perp T \mid X$ fails. No flexibility in estimating $\pi(x)$ or $\mu_t(x)$ can correct this.
- **Violations of consistency.** If the treatment version is ill-defined or SUTVA is violated, the potential outcomes framework is compromised.
- **Failure of positivity.** When $\pi(x) \approx 0$ or $\pi(x) \approx 1$, influence values become unstable and coverage degrades severely.
- **Ambiguity about the estimand.** Flexible estimation cannot resolve disagreement about what causal quantity is scientifically meaningful.
- **Invalid instrumental variable assumptions.** The exclusion restriction and exogeneity conditions (Chapter 7) are not testable; machine learning cannot substitute for subject-matter justification.
- **Fragile mediation assumptions.** Sequential ignorability and no interference between mediators (Chapter 8) must be argued on substantive grounds.

Machine learning improves the estimation of nuisance functions *within* a given identification strategy. It does not create identification where none exists.

> Remark
>
> Modern causal estimation is not merely machine learning plus a causal estimand: it is machine learning embedded inside a carefully constructed semiparametric procedure, whose validity rests on identification assumptions that no algorithm can verify.

12.14 Chapter Summary

Symbol	Meaning
η	Nuisance tuple (μ_0, μ_1, π)
$\phi(O;\, \psi, \eta)$	Generic orthogonal score; zero mean at truth
$\varphi_{\text{eff}}(O;\, \tau, \eta)$	Efficient influence function for the ATE Equation 12.4
$\mathcal{I}_k$	k-th fold; $k = 1, \ldots, K$
$\widehat{\eta}^{(-k)}$	Nuisance estimates trained without fold k
$\widehat{\tau}_{\text{DML}}$	Cross-fitted AIPW (DML) estimator Equation 12.11
$\widehat{\varphi}_i$	Out-of-fold estimated influence value for observation i
$\widehat{V}$	Variance estimator Equation 12.13

1. Flexible methods are attractive for nuisance estimation, but naive plug-in estimators can fail for two reasons: absent orthogonality, and same-sample nuisance fitting can create an additional empirical-process bias.
2. A score function is Neyman orthogonal if its Gateaux derivative with respect to the nuisance vanishes at the truth, so nuisance estimation error enters only at second order.
3. The efficient influence function for the ATE Equation 12.4 is orthogonal, doubly robust, and efficiency-achieving. It serves as the orthogonal score throughout this chapter.
4. Even with an orthogonal score, reusing the same data can induce overfitting bias, especially when the nuisance class falls outside classical fixed Donsker regimes.
5. Sample splitting removes this bias but wastes data. Cross-fitting recovers efficiency by rotating training and evaluation roles across K folds.
6. The cross-fitted AIPW estimator $\widehat{\tau}_{\text{DML}}$ is the standard DML estimator for the ATE.
7. Under the product-rate condition Equation 12.12, orthogonality and cross-fitting yield asymptotic linearity, root-n consistency, and asymptotic normality, with asymptotic variance equal to the

semiparametric efficiency bound (**?@thm-asymp**). A sufficient symmetric condition is that each nuisance estimator converges at rate $o_p(n^{-1/4})$.

8. The asymptotic variance is estimated consistently from the empirical variance of the out-of-fold estimated influence values.
9. Machine learning improves nuisance estimation, but it does not resolve identification problems: unmeasured confounding, positivity failures, and untestable structural assumptions remain the researcher's responsibility.

12.15 Problems

1. Verifying orthogonality.

(a) Write out the efficient influence function $\varphi_{\text{eff}}(O; \tau, \eta)$ for the ATE and compute its expectation. Confirm it equals zero at the truth.
(b) Consider a perturbation $\pi \mapsto \pi + r \cdot h$ for a bounded function $h(x)$. Differentiate $\mathbb{E}\{\varphi_{\text{eff}}(O; \tau, \mu_0, \mu_1, \pi + rh)\}$ with respect to r and evaluate at $r = 0$. Show the derivative is zero, confirming Neyman orthogonality in the propensity score direction.
(c) Repeat the calculation for a perturbation in μ_1.

2. First-order bias of the plug-in estimator. Suppose $\hat{\mu}_1 = \mu_1 + \delta$ for a deterministic function $\delta(x)$, and $\hat{\mu}_0 = \mu_0$, $\hat{\pi} = \pi$.

(a) Show that the bias of the prediction estimator $\hat{\tau}_{\text{pred}} = n^{-1} \sum_i \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}$ is $\mathbb{E}\{\delta(X)\}$.
(b) Show that the population bias of the plug-in AIPW estimator (using $\hat{\mu}_1 = \mu_1 + \delta$ with the true π and μ_0) is zero. Explain which property of the AIPW score is responsible. Does this mean the plug-in AIPW estimator is asymptotically unbiased when $\hat{\mu}_1$ is estimated flexibly from the same sample used to evaluate the score? Explain why or why not.

3. Cross-fitting with $K = 2$. Partition a sample of $n = 200$ observations into two halves $\mathcal{I}_1$ and $\mathcal{I}_2$.

(a) Write down the cross-fitted moment equation Equation 12.10 explicitly for $K = 2$.
(b) Explain why the contribution from $\mathcal{I}_1$ uses nuisance estimates trained on $\mathcal{I}_2$ and vice versa.
(c) Compare this procedure to single sample splitting with $\mathcal{I}_1$ as the training fold. What is gained and what is lost?

4. The product-rate condition. Suppose $\|\hat{\pi} - \pi\|_{L_2} = o_p(n^{-\alpha})$ and $\|\hat{\mu}_t - \mu_t\|_{L_2} = o_p(n^{-\beta})$ for $\alpha, \beta > 0$.

(a) State the condition on α and β under which the product-rate condition Equation 12.12 holds.
(b) Show that $\alpha = \beta = 1/4$ satisfies your condition.
(c) Does $\alpha = 1/2$, $\beta = 0$ satisfy it? What does this say about the case where the propensity score is estimated parametrically but the outcome model converges only at an unspecified rate?

5. Variance estimation. Using the explicit formula Equation 12.11, write out the estimated influence value $\hat{\varphi}_i$ for a generic observation $i \in \mathcal{I}_k$. Show that $n^{-1} \sum_i \hat{\varphi}_i = 0$ exactly when $\hat{\tau}_{\text{DML}}$ solves the cross-fitted moment equation, and explain why the centering in Equation 12.13 is numerically inconsequential.

6. What machine learning cannot do. For each of the following scenarios, identify the identification assumption that is violated and explain why flexible nuisance estimation cannot remedy the problem.

(a) A study of job-training effects on earnings omits pre-program earnings, a strong predictor of both program participation and subsequent earnings.
(b) A study of a medical treatment estimates the propensity score accurately, but the treatment was assigned in clusters (hospital wards) and outcomes may be correlated within clusters in a way that depends on the fraction of patients treated.
(c) An instrument is used to estimate the effect of education on wages, but the instrument also directly affects wages through a channel not related to education.

Chapter 13

Estimation under Instrumental Variables

Learning Objectives: Core Material

By the end of the core sections, students should be able to:

1. Construct the Wald estimator as the sample analog of the Wald estimand and interpret its numerator and denominator as the reduced form and first stage. Extend this to the IV regression estimator with covariates as the direct sample analog of $\widehat{\mathrm{Cov}}(\tilde{Z}, Y)/\widehat{\mathrm{Cov}}(\tilde{Z}, T)$, where $\tilde{Z}$ is the linear-projection residual of Z on the intercept and X.
2. Distinguish the structural form of a linear SEM from its reduced form; derive the reduced form by solving out endogenous variables; define the reduced form regression estimator; and interpret $\hat{\phi}_{\mathrm{RF}}$ as the intent-to-treat (ITT) effect when Z is randomly assigned.
3. Derive the 2SLS estimator from first principles via the constrained normal equations argument, and prove its numerical equivalence with the IV regression estimator.
4. Interpret 2SLS as a method-of-moments estimator based on the IV orthogonality condition, connect it to the estimating-equation framework of Chapter 10, and extend to GMM for overidentified models.
5. State the asymptotic distribution of the GMM estimator, specialize the sandwich variance formula to the exactly identified and efficient GMM cases, and construct cluster-robust standard errors.
6. Explain the weak-instrument problem and its consequences for finite-sample bias and inferential reliability.

Learning Objectives: Advanced Enrichment (Sections Section 13.8–Section 13.9)

Students who complete the advanced sections should additionally be able to:

1. Describe the GEL framework as a one-step alternative to efficient GMM; state the saddle-point formulation; identify empirical likelihood, exponential tilting, and the continuous updating estimator as special cases; and interpret the Legendre–Fenchel duality as observation reweighting.
2. Describe the control function approach, explain its equivalence to 2SLS in the linear model, and state the conditions under which $V = F_{T|Z,X}(T \mid Z, X)$ serves as a control variable in the nonlinear triangular model of Imbens and Newey (2009).

13.1 From Identification to Estimation

Chapter 7 established what IV identifies and under what assumptions. This chapter shows how to estimate that target from finite data using sample analogs of the same orthogonality restrictions.

Chapter 7 showed that when an instrument Z satisfies relevance, exogeneity, and exclusion, the residualized-

covariance ratio $\mathrm{Cov}(\tilde{Z}, Y)/\mathrm{Cov}(\tilde{Z}, T)$ is a well-defined function of observable quantities but does not identify any specific causal parameter without a fourth structural assumption. This chapter works within the linear constant-effect model:

$$Y = \alpha + \beta T + \gamma^\top X + \varepsilon, \tag{13.1}$$

$$T = a_T + \pi Z + \delta^\top X + \eta, \tag{13.2}$$

where the fourth structural assumption is the constant-effect restriction: $Y_i(1) - Y_i(0) = \beta$ for all i. Under the three core assumptions plus this restriction, the structural coefficient β is identified by the Wald formula:

$$\beta = \frac{\mathrm{Cov}(\tilde{Z},\, Y)}{\mathrm{Cov}(\tilde{Z},\, T)}.$$

This chapter addresses the estimation problem: how to recover β from a finite sample. The answer is not a single formula but a family of estimators — the Wald estimator, two-stage least squares (2SLS), and the generalized method of moments (GMM) — each of which is a sample analog of the same underlying orthogonality restriction.

Chapter structure. Core material (Section 13.2–Section 13.7) develops the mainstream IV estimators and their asymptotic theory. Advanced enrichment (Section 13.8–Section 13.9) introduces GEL and the control function approach.

13.2 The Wald Estimator and the IV Regression Estimator

13.2.1 The Wald Estimator

We begin with the simplest setting: a binary instrument $Z \in \{0, 1\}$, scalar treatment T, scalar outcome Y, and no additional covariates X.

Definition: Wald Estimator

Let $n_z = \sum_{i=1}^n \mathbf{1}(Z_i = z)$ and define within-group sample means $\bar{Y}_z = n_z^{-1} \sum_{i:\, Z_i = z} Y_i$, $\bar{T}_z = n_z^{-1} \sum_{i:\, Z_i = z} T_i$. The **Wald estimator** is:

$$\hat{\beta}_{\text{Wald}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{T}_1 - \bar{T}_0}. \tag{13.3}$$

The numerator $\bar{Y}_1 - \bar{Y}_0$ estimates the *reduced form*: the total effect of the instrument on the outcome. The denominator $\bar{T}_1 - \bar{T}_0$ estimates the *first stage*: the effect of the instrument on treatment uptake. Their ratio recovers the effect of treatment on the outcome by attributing all of the instrument's effect on Y to the path $Z \to T \to Y$ — valid precisely because the exclusion restriction rules out any direct path $Z \to Y$.

Remark: Consistency by the Continuous Mapping Theorem

The within-group means are consistent for $\mathbb{E}[Y \mid Z{=}z]$ and $\mathbb{E}[T \mid Z{=}z]$ by the LLN. Consistency of $\hat{\beta}_{\text{Wald}}$ for β then follows from the continuous mapping theorem, provided the denominator $\mathbb{E}[T \mid Z{=}1] - \mathbb{E}[T \mid Z{=}0] \neq 0$ (the relevance condition).

Remark: What the Wald Estimator Targets under Heterogeneous Effects

Binary treatment. For binary T, Chapter 7 showed that under the additional monotonicity assumption ($T_i(1) \geq T_i(0)$ for all i), the Wald estimand identifies the LATE for compliers. The Wald estimator consistently estimates the LATE, not the population ATE.
Continuous treatment. For continuous T with binary Z, the IV estimand identifies a weighted average of marginal causal effects $\partial Y_i(t)/\partial t$ across the treatment distribution, with weights determined by the instrument-induced shift in T. "The LATE" is no longer the right name for what is estimated.

13.2.2 The IV Regression Estimator with Covariates

When observed covariates $X \in \mathbb{R}^p$ are present, the Wald estimator is no longer applicable. We derive the general IV estimator from the estimating-equation principle.

Notation. Stack observations into $\mathbf{Y}, \mathbf{T}, \mathbf{Z} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times p}$. Let $\bar{\mathbf{X}} = [\mathbf{1}_n, \mathbf{X}]$ and define the annihilator matrix:

$$M_X = I_n - \bar{\mathbf{X}}(\bar{\mathbf{X}}^\top \bar{\mathbf{X}})^{-1} \bar{\mathbf{X}}^\top, \tag{13.4}$$

the orthogonal projection onto the orthogonal complement of $\mathrm{col}(\bar{\mathbf{X}})$.

Derivation via constrained normal equations. The structural model imposes:

$$\mathbb{E}\left[\begin{pmatrix} 1 \\ X_i \\ Z_i \end{pmatrix} \varepsilon_i\right] = 0, \qquad \mathbb{E}[T_i\, \varepsilon_i] \neq 0. \tag{13.5}$$

Consider the sample normal equations one would obtain by running OLS of $\mathbf{Y}$ on $(\mathbf{T}, \bar{\mathbf{X}}, \mathbf{Z})$, with residual $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{T}\beta - \bar{\mathbf{X}}\bar{\gamma}$:

$$\mathbf{T}^\top \hat{\mathbf{e}} = \mathbf{0}, \quad \text{(NE-T)} \quad \bar{\mathbf{X}}^\top \hat{\mathbf{e}} = \mathbf{0}, \quad \text{(NE-X)} \quad \mathbf{Z}^\top \hat{\mathbf{e}} = \mathbf{0}. \quad \text{(NE-Z)}$$

At the true β, (NE-T) fails because $\mathbb{E}[T_i \varepsilon_i] \neq 0$. We *drop* the invalid (NE-T) and solve the $(p+2)$-dimensional system (NE-X), (NE-Z). From (NE-X): $\hat{\bar{\gamma}} = (\bar{\mathbf{X}}^\top \bar{\mathbf{X}})^{-1}(\bar{\mathbf{X}}^\top \mathbf{Y} - \bar{\mathbf{X}}^\top \mathbf{T}\beta)$. Substituting into (NE-Z) and simplifying using M_X: $\mathbf{Z}^\top M_X \mathbf{T}\, \beta = \mathbf{Z}^\top M_X \mathbf{Y}$.

> Definition: IV Regression Estimator
>
> The **IV regression estimator** in the linear IV model with scalar instrument Z and covariates X is:
>
> $$\hat{\beta}_{\mathrm{IV}} = \frac{\mathbf{Z}^\top M_X \mathbf{Y}}{\mathbf{Z}^\top M_X \mathbf{T}} = \frac{\widehat{\mathrm{Cov}}(\tilde{Z}, Y)}{\widehat{\mathrm{Cov}}(\tilde{Z}, T)}, \tag{13.6}$$
>
> where $\tilde{Z}_i$, $\tilde{Y}_i$, $\tilde{T}_i$ are the within-X residuals (entries of $M_X\mathbf{Z}$, $M_X\mathbf{Y}$, $M_X\mathbf{T}$).

The Wald estimator is the special case $Z \in \{0, 1\}$, no covariates: $M_X = I_n - n^{-1}\mathbf{1}\mathbf{1}^\top$ is the centering matrix, and $\hat{\beta}_{\mathrm{IV}} = \hat{\beta}_{\mathrm{Wald}}$.

13.2.3 Structural Form, Reduced Form, and the Reduced Form Regression

The structural form. System Equation 13.1–Equation 13.2 is the *structural form*: each equation describes how one variable is determined by others, including endogenous variables on the right-hand side. The structural coefficient β has a causal interpretation, but T is correlated with ε, so OLS applied to the structural form is inconsistent.

The reduced form. The *reduced form* is obtained by solving the structural system so each endogenous variable is expressed purely as a function of exogenous variables Z and X. Substituting Equation 13.2 into Equation 13.1:

$$Y = \underbrace{(\alpha + \beta a_T)}_{\alpha_{\mathrm{rf}}} + \underbrace{\beta\pi}_{\phi} Z + \underbrace{(\beta\delta + \gamma)^\top}_{\gamma_{\mathrm{rf}}^\top} X + \underbrace{(\beta\eta + \varepsilon)}_{\nu}. \tag{13.7}$$

Writing compactly: $Y = \alpha_{\mathrm{rf}} + \phi Z + \gamma_{\mathrm{rf}}^\top X + \nu$, where $\phi = \beta\pi$, $\nu = \beta\eta + \varepsilon$. Both right-hand-side variables (Z and X) are exogenous, so OLS is consistent for ϕ. Neither β nor π is separately identified from the reduced form alone.

Definition: Reduced Form Regression Estimator

The **reduced form regression estimator** is OLS of ϕ in the reduced form equation Equation 13.7:

$$\hat{\phi}_{\text{RF}} = \frac{\widehat{\text{Cov}}(\tilde{Z}, Y)}{\widehat{\text{Var}}(\tilde{Z})}, \tag{13.8}$$

consistent for $\phi = \beta\pi$ under $\mathbb{E}[\nu \mid Z, X] = 0$.

Relationship to the first stage and IV estimator. The first-stage regression estimator is $\hat{\pi}_{\text{FS}} = \widehat{\text{Cov}}(\tilde{Z}, T)/\widehat{\text{Var}}(\tilde{Z})$. Since $\phi = \beta\pi$, the sample analog gives:

$$\hat{\beta}_{\text{IV}} = \frac{\hat{\phi}_{\text{RF}}}{\hat{\pi}_{\text{FS}}}, \tag{13.9}$$

reproducing Equation 13.6. The reduced form delivers the instrument's total effect on the outcome; the first stage scales it by the instrument's effect on treatment; the ratio recovers the structural parameter.

Intent-to-treat interpretation. When Z is randomly assigned, $\hat{\phi}_{\text{RF}}$ estimates the *intent-to-treat* (ITT) effect: the average effect on Y of being assigned $Z = 1$ rather than $Z = 0$, regardless of actual treatment uptake. The ITT requires only exogeneity of Z, not the exclusion restriction, and is often of direct policy interest.

Remark: Reporting All Three Quantities

In applied work it is standard practice to report the first stage, reduced form, and IV (or 2SLS) estimates side by side. The reduced form and first stage each have a transparent OLS interpretation and can be assessed independently before the ratio is formed. The IV estimate carries no identifying content beyond what the reduced form and first stage together contain.

13.3 Two-Stage Least Squares

Two-stage least squares (2SLS) extends the Wald estimator to settings with continuous or multi-valued instruments, multiple instruments, and observed covariates.

Stage 1: The First-Stage Regression. Regress T on Z and X by OLS, obtaining fitted values $\hat{T}_i = \hat{a}_T + \hat{\pi}^\top Z_i + \hat{\delta}^\top X_i$. In matrix form, $\hat{\mathbf{T}} = P_W \mathbf{T}$ where P_W is the projection onto the column space of the instrument design matrix $\mathbf{W}$. The fitted values $\hat{T}_i$ isolate the component of treatment variation spanned by $(1, Z, X)$ — the variation that is exogenous under IV validity.

Stage 2: The Second-Stage Regression. Regress Y on $\hat{T}$ and X by OLS.

Definition: 2SLS Estimator

The **two-stage least squares estimator** $\hat{\beta}_{\text{2SLS}}$ is the OLS coefficient on $\hat{T}$ in the second-stage regression. In the single-instrument case ($q = 1$):

$$\hat{\beta}_{\text{2SLS}} = \frac{\sum_{i=1}^n \tilde{Z}_i Y_i}{\sum_{i=1}^n \tilde{Z}_i T_i} = \frac{\widehat{\text{Cov}}(\tilde{Z}, Y)}{\widehat{\text{Cov}}(\tilde{Z}, T)}. \tag{13.10}$$

The component of T orthogonal to (Z, X) — the residual $\hat{\eta}_i = T_i - \hat{T}_i$, correlated with ε_i when T is endogenous — is dropped before the causal coefficient is estimated.

Standard Errors from the Second-Stage OLS Are Incorrect

A common error is to report the standard errors from the second-stage OLS regression directly. Those standard errors use $\hat{T}$ rather than T and residuals that do not equal the structural errors ε_i,

so they are invalid. Correct inference requires the sandwich variance formula from Section 13.6, or software that implements 2SLS natively.

13.4 Equivalence of the IV Regression Estimator and 2SLS

Theorem: IV Regression–2SLS Equivalence

In the linear IV model with a scalar instrument Z and covariates $X \in \mathbb{R}^p$, the IV regression estimator and the 2SLS estimator coincide: $\hat{\beta}_{\text{2SLS}} = \hat{\beta}_{\text{IV}}$.

Proof

First-stage projection. The first-stage fitted value is $\hat{\mathbf{T}} = P_{[\bar{\mathbf{X}},\mathbf{Z}]}\mathbf{T}$. By the Frisch–Waugh–Lovell theorem:

$$M_X\hat{\mathbf{T}} = M_X\mathbf{Z}(\mathbf{Z}^\top M_X\mathbf{Z})^{-1}\mathbf{Z}^\top M_X\mathbf{T}.$$

2SLS second-stage closed form. Applying FWL to the second-stage regression of $\mathbf{Y}$ on $(\hat{\mathbf{T}}, \bar{\mathbf{X}})$:

$$\hat{\beta}_{\text{2SLS}} = \frac{\hat{\mathbf{T}}^\top M_X\mathbf{Y}}{\hat{\mathbf{T}}^\top M_X\hat{\mathbf{T}}} = \frac{\mathbf{T}^\top M_X\mathbf{Z}(\mathbf{Z}^\top M_X\mathbf{Z})^{-1}\mathbf{Z}^\top M_X\mathbf{Y}}{\mathbf{T}^\top M_X\mathbf{Z}(\mathbf{Z}^\top M_X\mathbf{Z})^{-1}\mathbf{Z}^\top M_X\mathbf{T}}.$$

Under exact identification ($q = 1$), the scalars $\mathbf{T}^\top M_X\mathbf{Z}$ and $(\mathbf{Z}^\top M_X\mathbf{Z})^{-1}$ cancel from numerator and denominator, leaving $\hat{\beta}_{\text{2SLS}} = \mathbf{Z}^\top M_X\mathbf{Y}/\mathbf{Z}^\top M_X\mathbf{T} = \hat{\beta}_{\text{IV}}$. □

Remark: What the Proof Reveals

In OLS, the normal equation for each regressor forces the residual to be orthogonal to that regressor. When T is endogenous, its normal equation is corrupted: $\mathbb{E}[\mathbf{T}^\top\varepsilon] \neq 0$. IV simply replaces this corrupted equation with the instrument equation $\mathbf{Z}^\top\hat{\mathbf{e}} = 0$, which is valid because Z is exogenous. The IV estimator is OLS on the structural model but with one invalid normal equation swapped for a valid one. The 2SLS derivation reaches the same formula by a projection argument.

In the scalar-instrument linear model, Wald, IV regression, and 2SLS are not competing methods; they are different representations of the same sample analog of the identification formula.

13.5 The Moment-Condition View and GMM

13.5.1 2SLS as a Method-of-Moments Estimator

Stack the constant, covariates, and instrument into $W = (1, X^\top, Z^\top)^\top$. The IV moment condition is $\mathbb{E}[W\,(Y - \alpha - \beta T - \gamma^\top X)] = 0$. Setting $\theta = (\alpha, \beta, \gamma^\top)^\top$ and defining $U(O;\theta) = W\,(Y - D^\top\theta)$ where $D = (1, T, X^\top)^\top$, the identifying condition is $\mathbb{E}\{U(O;\theta_0)\} = 0$ — exactly an estimating equation in the sense of Chapter 10.

Theorem: 2SLS as a Method-of-Moments Estimator

In the exactly identified linear IV model ($q = 1$), the 2SLS estimator $\hat{\theta}_{\text{2SLS}}$ is the unique solution to the sample moment system:

$$\frac{1}{n}\sum_{i=1}^{n} W_i\,(Y_i - \hat{\alpha} - \hat{\beta}_{\text{2SLS}}\,T_i - \hat{\gamma}^\top X_i) = 0, \qquad W_i = (1,\, X_i^\top,\, Z_i)^\top.$$

13.5.2 Overidentification and GMM

When $q > 1$ instruments are available, the model is *overidentified*: more moment conditions than parameters. The system $\mathbb{P}_n U(O;\theta) = 0$ is generically overdetermined and has no exact solution.

The *generalized method of moments* (GMM) minimizes a weighted quadratic form in the sample moments:

$$\hat{\theta}_{\text{GMM}} = \arg\min_{\theta} \left[\mathbb{P}_n U(O;\theta)\right]^\top \hat{\Omega}_n \left[\mathbb{P}_n U(O;\theta)\right].$$

Different choices of $\hat{\Omega}_n$ yield different estimators: $\hat{\Omega}_n = (n^{-1}\sum_i W_i W_i^\top)^{-1}$ yields 2SLS; $\hat{\Omega}_n = [\mathbb{P}_n U(O;\hat{\theta}) U(O;\hat{\theta})^\top]^{-1}$ yields the *efficient GMM estimator*. Under homoskedasticity, efficient GMM and 2SLS coincide. Under heteroskedasticity, efficient GMM is weakly (and generically strictly) more efficient.

> Remark: Overidentification as a Testable Restriction
>
> An overidentified model imposes more moment conditions than needed for point identification. The Sargan–Hansen J-statistic (Sargan 1958) tests the joint null that all moment conditions hold: under H_0, $J \xrightarrow{d} \chi^2_{q-1}$. A rejection indicates that the full set of maintained moment conditions is incompatible with the data. The test does not identify which assumption failed.

> Example: GMM with Two Instruments
>
> Consider $Y = \beta T + \varepsilon$ (after demeaning), with two instruments Z_1, Z_2 giving moment conditions $\mathbb{E}[Z_j \varepsilon] = 0$, $j = 1, 2$.
> Sample covariances: $\widehat{\text{Cov}}(Z_1, Y) = 0.40$, $\widehat{\text{Cov}}(Z_1, T) = 0.50$, $\widehat{\text{Cov}}(Z_2, Y) = 0.60$, $\widehat{\text{Cov}}(Z_2, T) = 0.80$.
> Just-identified IV estimates: $\hat{\beta}^{(1)} = 0.40/0.50 = 0.80$, $\hat{\beta}^{(2)} = 0.60/0.80 = 0.75$.
> **GMM with identity weighting ($\hat{\Omega} = I_2$).** With $a = (0.50, 0.80)^\top$ and $c = (0.40, 0.60)^\top$, the sample moment vector is $\hat{U}(\beta) = c - a\beta$. The first-order condition $a^\top(c - a\beta) = 0$ gives:
>
> $$\hat{\beta}_{\text{GMM}} = \frac{a^\top c}{a^\top a} = \frac{(0.50)(0.40) + (0.80)(0.60)}{(0.50)^2 + (0.80)^2} = \frac{0.68}{0.89} \approx 0.764.$$
>
> A precision-weighted average lying between 0.75 and 0.80, with more weight on Z_2 (larger first-stage covariance).
> **Sargan–Hansen J-test.** Residual moments at $\hat{\beta} = 0.764$: $\hat{U}_1 = 0.018$, $\hat{U}_2 = -0.011$. The J-statistic $J = n\,\hat{U}(\hat{\beta})^\top \hat{\Sigma}^{-1} \hat{U}(\hat{\beta})$ has a χ^2_1 null distribution. A failure to reject is consistent with both instruments satisfying the exclusion restriction.

13.6 Asymptotic Theory of the GMM Estimator

13.6.1 Setup and Notation

Stack the structural regressors and instruments:

$$D_i = (1,\, T_i,\, X_i^\top)^\top \in \mathbb{R}^k, \qquad W_i = (1,\, X_i^\top,\, Z_i^\top)^\top \in \mathbb{R}^m,$$

where $k = p+2$ and $m = p+1+q$. The IV moment condition is $\mathbb{E}[U(O_i;\theta_0)] = 0$, $U(O_i;\theta) = W_i(Y_i - D_i^\top\theta)$. Define the sensitivity matrix $A = \mathbb{E}[W_i D_i^\top] \in \mathbb{R}^{m\times k}$ and moment variance matrix $\Sigma = \mathbb{E}[\varepsilon_i^2\, W_i W_i^\top]$.

13.6.2 Asymptotic Distribution

> Theorem: Asymptotic Distribution of the GMM Estimator
>
> Suppose the IV moment condition holds, O_i are i.i.d. with finite fourth moments, A has full column rank, and Σ is positive definite. Then:
>
> $$\sqrt{n}\,(\hat{\theta}_{\text{GMM}} - \theta_0) \xrightarrow{d} N(0,\, V_{\text{GMM}}(\Omega)),$$

where the **sandwich variance** is:

$$V_{\text{GMM}}(\Omega) = (A^\top \Omega A)^{-1} A^\top \Omega \Sigma \Omega A (A^\top \Omega A)^{-1}. \tag{13.11}$$

The formula is an instance of the general estimating-equation theory from Chapter 10: the sensitivity matrix A plays the role of $-\mathbb{E}[\partial U/\partial \theta^\top]$ and the moment variance Σ plays the role of $\mathbb{E}[UU^\top]$.

13.6.3 Two Important Special Cases

Exactly identified case ($q = 1$, $m = k$). A is square and invertible. All weighting matrices yield the same estimator. The sandwich variance simplifies to:

$$V_{\text{IV}} = A^{-1} \Sigma (A^\top)^{-1}. \tag{13.12}$$

In the scalar no-covariate case ($p = 0$), the block corresponding to β is:

$$V_\beta = \frac{\mathbb{E}[\varepsilon^2 \tilde{Z}^2]}{(\mathbb{E}[\tilde{Z} T])^2}, \qquad \tilde{Z} = Z - \mathbb{E}[Z].$$

Under homoskedasticity and first-stage relation $T = a_T + \pi Z + \eta$: $V_\beta = \sigma^2/(\pi^2 \operatorname{Var}(Z))$.

Efficient GMM. The asymptotic variance $V_{\text{GMM}}(\Omega)$ is minimized by $\Omega^* = \Sigma^{-1}$, giving:

$$V_{\text{eff}} = (A^\top \Sigma^{-1} A)^{-1}. \tag{13.13}$$

This is the semiparametric efficiency lower bound for IV estimation within the linear IV moment-restriction model $\mathbb{E}[W\varepsilon] = 0$.

> Remark: When 2SLS Is Efficient
>
> The 2SLS weighting matrix is $\Omega_{\text{2SLS}} = \mathbb{E}[W_i W_i^\top]^{-1}$. Under homoskedasticity, $\Sigma = \sigma^2 \mathbb{E}[WW^\top]$, so $\Sigma^{-1} \propto \Omega_{\text{2SLS}}$, and 2SLS achieves V_{eff}. Under heteroskedasticity, efficient GMM is weakly (and generically strictly) more efficient.

13.6.4 Consistent Variance Estimation

Let $\hat{\varepsilon}_i = Y_i - D_i^\top \hat{\theta}$ denote the structural residuals. Consistent estimators: $\hat{A} = n^{-1} \sum_i W_i D_i^\top$, $\hat{\Sigma} = n^{-1} \sum_i \hat{\varepsilon}_i^2 W_i W_i^\top$. The heteroskedasticity-robust sandwich variance estimator is:

$$\hat{V}_{\text{GMM}} = (\hat{A}^\top \hat{\Omega}_n \hat{A})^{-1} \hat{A}^\top \hat{\Omega}_n \hat{\Sigma} \hat{\Omega}_n \hat{A} (\hat{A}^\top \hat{\Omega}_n \hat{A})^{-1}. \tag{13.14}$$

In the exactly identified case, Equation 13.14 reduces to $\hat{A}^{-1} \hat{\Sigma} (\hat{A}^\top)^{-1}$, independent of $\hat{\Omega}_n$. In applications, the default should be heteroskedasticity-robust standard errors; cluster-robust standard errors are required when observations within groups share unmodeled common shocks.

> Remark: Two-Step Efficient GMM
>
> Because the optimal weighting $\hat{\Omega}_n = \hat{\Sigma}^{-1}$ requires preliminary residuals $\hat{\varepsilon}_i$, efficient GMM is typically implemented in two steps: obtain a consistent first-step estimator (e.g., 2SLS) to form $\hat{\varepsilon}_i$, construct $\hat{\Sigma}$, and re-minimize the GMM criterion with $\hat{\Omega}_n = \hat{\Sigma}^{-1}$.

13.7 Weak Instruments and Inferential Fragility

When the first-stage relationship is weak, IV estimation becomes severely fragile. A weak instrument is not merely an efficiency problem: it makes the finite-sample distribution of IV estimators highly non-normal, magnifies bias toward OLS, and undermines conventional confidence intervals.

13.7.1 The Weak-Instrument Problem

The 2SLS closed form Equation 13.10 divides by the sample covariance $\widehat{\mathrm{Cov}}(\tilde{Z}, T)$. When the population first-stage coefficient π is near zero, the consequences are (Bound et al. 1995):

- **Finite-sample bias toward OLS.** As $\pi \to 0$, the 2SLS bias approaches the OLS bias rather than zero.
- **Non-Gaussian finite-sample distribution.** The distribution of $\hat{\beta}_{\mathrm{2SLS}}$ can be highly skewed or heavy-tailed, rendering the $N(0, V_{\mathrm{2SLS}})$ approximation unreliable.
- **Size distortion.** Wald-type confidence intervals can severely undercover the true parameter.

13.7.2 Diagnostic: The First-Stage F-Statistic

The most widely used diagnostic is the F-statistic from the first-stage regression, testing the joint significance of Z after partialling out X. Staiger and Stock (1997) argued informally for $F \geq 10$ as adequate instrument strength; Stock and Yogo (2005) provided formal critical values. **A large first-stage F supports relevance, but says nothing directly about exogeneity or exclusion.** The first-stage F-statistic is a relevance diagnostic, not a certificate of instrument validity.

> The $F > 10$ Rule of Thumb
>
> The threshold $F_{\text{first stage}} > 10$ is a widely cited but coarse diagnostic. It targets finite-sample bias relative to OLS, not validity of the exclusion restriction. A more reliable diagnostic is the effective F-statistic of Olea and Pflueger (2013), which remains valid under heteroskedasticity and within-cluster correlation.

13.7.3 Weak-Instrument-Robust Inference

When first-stage strength is uncertain, alternative inferential procedures with size guarantees are needed.

The **Anderson–Rubin (AR) test** (Anderson and Rubin, 1949) inverts the question: it tests whether a hypothesized value β_0 is consistent with the IV moment condition. Substituting β_0 into the structural model gives $Y - \beta_0 T = \alpha + \gamma^\top X + (\varepsilon + (\beta - \beta_0)T)$. If $\beta_0 = \beta$, the composite error is uncorrelated with Z by instrument validity. The AR test regresses $Y - \beta_0 T$ on Z and X and tests the null that the coefficient on Z is zero. Under the classical homoskedastic Gaussian linear model the F-statistic has an *exact* finite-sample F-distribution regardless of instrument strength. Inverting this test yields a confidence set valid whether or not the instrument is weak.

The **conditional likelihood ratio (CLR) test** of Moreira (2003) extends the AR idea more efficiently to the multiple-instrument case.

> Remark: Two Distinct IV Concerns
>
> The weak-instrument problem is a *statistical* concern: it questions the quality of estimation given a valid identification strategy. It is distinct from the *interpretive* concern of Chapter 7 — that even a perfectly strong instrument identifies only the LATE for compliers, not the population ATE. Any complete IV analysis must address both.

13.8 Generalized Empirical Likelihood

> Advanced Topic
>
> This section introduces GEL as a one-step alternative to efficient GMM. The main takeaway is that every GEL estimator shares the same first-order asymptotic efficiency as efficient GMM, while GEL can have better higher-order finite-sample properties in overidentified models. The duality formulation and special-cases table may be treated as optional reading.

GEL uses the same IV moment restrictions but incorporates them through an implied reweighting of the sample rather than through a separately estimated covariance weight matrix. Section Section 13.6 showed

that efficient GMM requires a two-step procedure; this two-step structure introduces finite-sample bias (Newey and Smith 2004). GEL provides a one-step alternative.

13.8.1 The GEL Estimator

Let $U_i(\theta) = W_i(Y_i - D_i^\top \theta)$ and $\bar{U}(\theta) = n^{-1}\sum_i U_i(\theta)$. GEL introduces a strictly convex function $G\colon \mathcal{V} \to \mathbb{R}$ (open interval $\mathcal{V}$ containing zero) and solves the saddle-point problem:

$$\hat{\theta}_{\text{GEL}} = \arg\min_{\theta} \sup_{\lambda \in \Lambda_n(\theta)} \frac{1}{n}\sum_{i=1}^{n}[-G(\lambda^\top U_i(\theta))], \tag{13.15}$$

where $\lambda \in \mathbb{R}^m$ is an auxiliary dual variable. G is normalized: $G(0) = 0$, $g(0) = 1$, $g'(0) = 1$ where $g = G'$.

Definition: Special Cases of GEL

1. **Empirical likelihood (EL):** $G(v) = -\log(1 - v)$, $\mathcal{V} = (-\infty, 1)$.
2. **Exponential tilting (ET):** $G(v) = e^v - 1$, $\mathcal{V} = \mathbb{R}$.
3. **Continuous updating estimator (CUE):** $G(v) = v + v^2/2$, $\mathcal{V} = \mathbb{R}$. The GEL saddle-point problem reduces to simultaneous minimization of the GMM criterion with the weighting matrix $\hat{\Omega}(\theta) = [n^{-1}\sum_i U_i(\theta)U_i(\theta)^\top]^{-1}$ continuously updated at the current θ.

13.8.2 The Convex-Conjugate Duality (Optional)

The GEL saddle-point problem Equation 13.15 is the Lagrangian dual of a *minimum-discrepancy* (MD) primal problem that re-weights observations. Define the Legendre–Fenchel conjugate of G:

$$F(\omega) = \sup_{v \in \mathcal{V}} [\omega v - G(v)], \tag{13.16}$$

a strictly convex function with $F(1) = 0$ (by the normalization). The MD estimator minimizes a convex divergence between observation weights ω_i and the reference $\omega_i = 1$, subject to $\sum_i \omega_i U_i(\theta) = 0$.

Theorem: GEL–MD Duality [@neweysmith2004; @ragusa2011]

The GEL problem Equation 13.15 is the Lagrangian dual of the MD problem: their first-order conditions coincide, so $\hat{\theta}_{\text{GEL}} = \hat{\theta}_{\text{MD}}$. The dual variable $\hat{\lambda}$ and primal weights $\hat{\omega}_i$ are related by $\hat{\omega}_i = g(\hat{\lambda}^\top U_i(\hat{\theta}))$.

The GEL special cases correspond to different divergences: EL uses reverse KL; ET uses forward KL; CUE uses a quadratic distance.

Remark: Re-weighting Observations vs. Re-weighting Moments

GMM re-weights the *moment conditions* via the matrix $\hat{\Omega}$. MD/GEL instead re-weights the *observations*, finding the empirical distribution closest to uniform in the F-divergence sense consistent with the moment restrictions. The two approaches are asymptotically equivalent to first order.

13.8.3 Asymptotic Properties and Comparison with GMM

Theorem: First-Order Asymptotic Equivalence [@neweysmith2004]

Under regularity conditions, every GEL estimator achieves the efficient GMM variance:

$$\sqrt{n}\,(\hat{\theta}_{\text{GEL}} - \theta_0) \xrightarrow{d} N(0,\ (A^\top \Sigma^{-1} A)^{-1})\,.$$

Theorem: GEL Overidentification Test (Wilks' Theorem) [@neweysmith2004]

Under the same conditions, the GEL profile divergence satisfies:

$$T_{\text{GEL}} \equiv 2\sum_{i=1}^{n} F(\hat{\omega}_i) = -2\sum_{i=1}^{n} G(\hat{\lambda}^\top U_i(\hat{\theta}_{\text{GEL}})) \xrightarrow{d} \chi^2_{m-k}. \tag{13.17}$$

For EL specifically, $T_{\text{EL}} = -2\sum_i \log(n\hat{\pi}_i)$, the empirical likelihood ratio statistic. To higher order, GEL estimators have smaller bias than two-step GMM: GEL eliminates the bias from estimating the Jacobian; EL additionally eliminates bias from estimating the weighting matrix Σ.

13.9 The Control Function Approach

The control-function approach offers an alternative route to handling endogeneity: instead of projecting treatment onto instruments, it augments the outcome model with a control variable that absorbs the endogenous component of treatment selection.

2SLS achieves identification by replacing the endogenous regressor T with its exogenous projection $\hat{T}$. The *control function* approach adds the first-stage residual to the outcome regression as an explicit control for the endogenous variation, rather than removing it from the treatment variable.

13.9.1 Linear Model and Equivalence to 2SLS

The linear control-function representation requires the **linear control-function assumption**:

$$\mathbb{E}[\varepsilon \mid \eta, Z, X] = \rho\,\eta, \qquad \rho = \frac{\text{Cov}(\varepsilon, \eta)}{\text{Var}(\eta)}. \tag{13.18}$$

This holds under joint normality of (ε, η) given (Z, X). Defining $\xi = \varepsilon - \rho\eta$, assumption Equation 13.18 is equivalent to $\mathbb{E}[\xi \mid \eta, Z, X] = 0$. Substituting into the outcome equation: $Y = \alpha + \beta T + \gamma^\top X + \rho\eta + \xi$. Including η as an additional regressor renders T exogenous in the augmented regression.

Control Function Estimator (Linear Case)

1. **First stage.** Regress T on Z and X by OLS; obtain residuals $\hat{\eta}_i = T_i - \hat{\pi}^\top Z_i - \hat{\delta}^\top X_i$.
2. **Second stage.** Regress Y on T, X, and $\hat{\eta}$ by OLS; the coefficient on T is the control function estimator $\hat{\beta}_{\text{CF}}$.

Theorem: Equivalence of 2SLS and the Control Function Estimator

In the linear IV model, $\hat{\beta}_{\text{CF}}$ from the augmented OLS regression of Y on $(T, X, \hat{\eta})$ is numerically identical to $\hat{\beta}_{\text{2SLS}}$.

Proof

The Frisch–Waugh–Lovell theorem states that the coefficient on T in the OLS regression of Y on $(T, X, \hat{\eta})$ equals the coefficient on $\tilde{T}$ in the regression of $\tilde{Y}$ on $\tilde{T}$, where tildes denote residuals after projecting out $(X, \hat{\eta})$. Since $\hat{\eta} = T - \hat{T}$, projecting out $\hat{\eta}$ and X from T is equivalent to projecting out X and $T - \hat{T}$ from T, which leaves $\hat{T}$ after partialling out X. Thus $\tilde{T}$ is the residual from projecting $\hat{T}$ on X — the same residual used in the 2SLS second stage. □

Standard Errors in the Second Stage

The coefficient on T is identical to 2SLS, but the OLS standard errors are not. They ignore the sampling variation in $\hat{\eta}_i$ and are therefore incorrect. Correct inference requires the 2SLS sandwich formula, or a bootstrap that resamples both stages jointly.

13.9.2 Testing Endogeneity via Residual Inclusion

The control function representation yields a natural test of $H_0 : \rho = 0$ (exogeneity of T). The t-statistic on $\hat{\eta}$ in the augmented regression tests endogeneity. A rejection suggests endogeneity under the maintained instrument assumptions; it is not a stand-alone validation of the instrument, since the test takes instrument validity as given. This is the regression-based form of the Hausman (1978) endogeneity test.

13.9.3 Brief Note on Nonlinear Extensions

2SLS does not carry over. In a probit or Poisson outcome model, 2SLS is generally inconsistent: plugging in $\hat{T}$ breaks the nonlinear link function.

Related control-variable methods exist. Imbens and Newey (2009) showed that under *independence*, $(\varepsilon, \eta) \perp\!\!\!\perp Z \mid X$, and *scalar monotonicity*, $T = h(Z, X, \eta)$ strictly monotonic in a scalar η, the conditional CDF:

$$V = F_{T|Z,X}(T \mid Z, X) \tag{13.19}$$

is a valid control variable in the sense that $T \perp\!\!\!\perp \varepsilon \mid X, V$. Conditioning on (X, V) recovers structural variation in T, allowing identification of the average structural function. Note that X must be retained in the conditioning set; dropping it gives the stronger $T \perp\!\!\!\perp \varepsilon \mid V$, which fails whenever X has any direct effect on ε.

13.10 Chapter Summary

Symbol	Meaning
$\hat{\beta}_{\text{Wald}}$	Wald estimator: $(\bar{Y}_1 - \bar{Y}_0)/(\bar{T}_1 - \bar{T}_0)$
$\hat{\phi}_{\text{RF}}$	Reduced form regression estimator of $\phi = \beta\pi$
$\hat{\pi}_{\text{FS}}$	First-stage regression estimator of π
M_X	Annihilator matrix (within-X residuals)
$\hat{\beta}_{\text{IV}}$	IV regression estimator Equation 13.6
$\hat{\beta}_{\text{2SLS}}$	2SLS estimator Equation 13.10
$V_{\text{GMM}}(\Omega)$	Sandwich variance Equation 13.11
V_{eff}	Efficient GMM variance Equation 13.13
$\hat{V}_{\text{GMM}}$	Consistent variance estimator Equation 13.14
J-statistic	Sargan–Hansen overidentification test
$\hat{\theta}_{\text{GEL}}$	GEL estimator Equation 13.15

1. **Wald, IV regression, and 2SLS are one estimator.** The Wald estimator, IV regression estimator, and 2SLS are numerically identical in the single-instrument case: they are different representations of the same sample analog of the identification formula. 2SLS extends to multiple instruments; the Wald estimator is the further special case $Z \in \{0, 1\}$, X absent.
2. **Structural form vs. reduced form.** The structural form contains endogenous regressors; the reduced form expresses each endogenous variable as a function of exogenous variables. The reduced form coefficient $\phi = \beta\pi$ is estimable by OLS; β is recovered only via the ratio $\hat{\phi}_{\text{RF}}/\hat{\pi}_{\text{FS}}$.
3. **2SLS as moment-of-moments.** 2SLS is the method-of-moments estimator for the IV orthogonality condition $\mathbb{E}[W\varepsilon] = 0$, an instance of Chapter 10's estimating-equation framework.
4. **GMM and efficient weighting.** Efficient GMM achieves the semiparametric efficiency bound; 2SLS is efficient under homoskedasticity but not generally under heteroskedasticity. All standard errors should be heteroskedasticity-robust; cluster-robust when observations within groups share common shocks.
5. **Weak instruments.** Weak instruments cause finite-sample bias toward OLS, non-Gaussian distributions, and size distortion. The first-stage F-statistic is a relevance diagnostic, not a validity certificate. Anderson–Rubin confidence sets provide weak-instrument-robust inference.
6. **GEL.** GEL estimators achieve the efficient GMM variance in one step without a preliminary weighting step. To higher order, they have smaller bias than two-step GMM in overidentified models.
7. **Control function.** In the linear model, the control function approach is numerically equivalent to 2SLS; it provides a direct test of endogeneity via the t-statistic on $\hat{\eta}$. In nonlinear models, related

control-variable methods exist under the stronger independence and scalar-monotonicity conditions of Imbens and Newey (2009).

13.11 Problems

1. The Wald estimator as a ratio-of-moments estimator.

(a) Augment β with an intercept α and express the structural model as the solution to the two-dimensional moment condition $\mathbb{E}[(1, Z)^\top (Y - \alpha - \beta T)] = 0$. Verify exact identification and solve to recover $\beta = \Delta_Y / \Delta_T$.

(b) Using $\bar{Y}_z \xrightarrow{p} \mu_Y(z)$ and $\bar{T}_z \xrightarrow{p} \mu_T(z)$, prove $\hat{\beta}_{\text{Wald}} \xrightarrow{p} \beta$ via the continuous mapping theorem.

(c) Apply the delta method to $(\hat{\Delta}_Y, \hat{\Delta}_T)^\top$ to show $\sqrt{n}(\hat{\beta}_{\text{Wald}} - \beta) \xrightarrow{d} N(0, V)$ where $V = \Delta_T^{-2} \sum_{z \in \{0,1\}} \text{Var}(Y_i - \beta T_i \mid Z_i = z)/p_z$, and confirm this matches the IV variance formula Equation 13.12 in the scalar no-covariate case.

2. Matrix form of 2SLS and why second-stage standard errors are wrong.

Let $\mathbf{D} \in \mathbb{R}^{n \times k}$ be the full regressor matrix and $\mathbf{W} \in \mathbb{R}^{n \times m}$ the full instrument matrix. Let $P_{\mathbf{W}} = \mathbf{W}(\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top$.

(a) Show the 2SLS estimator can be written as $\hat{\theta}_{\text{2SLS}} = (\mathbf{D}^\top P_{\mathbf{W}} \mathbf{D})^{-1} \mathbf{D}^\top P_{\mathbf{W}} \mathbf{Y}$.

(b) In the single-instrument, no-covariate case, verify from the matrix formula that $\hat{\beta}_{\text{2SLS}} = \hat{\beta}_{\text{IV}}$.

(c) The second-stage OLS uses $\hat{\mathbf{D}}$ in place of $\mathbf{D}$. Let $\hat{\varepsilon}_{\text{2nd}} = \mathbf{Y} - \hat{\mathbf{D}} \hat{\theta}_{\text{2SLS}}$. Show $\hat{\varepsilon}_{\text{2nd}} \neq \mathbf{Y} - \mathbf{D} \hat{\theta}_{\text{2SLS}}$ in general. Explain why this discrepancy makes the second-stage OLS standard errors invalid, and identify the correct residuals for the sandwich variance Equation 13.14.

3. Efficiency of GMM and the Sargan–Hansen J-statistic.

(a) Prove $V_{\text{GMM}}(\Omega) \succeq V_{\text{eff}}$ for every positive-definite Ω, where $V_{\text{eff}} = (A^\top \Sigma^{-1} A)^{-1}$. (*Hint:* factor $V_{\text{GMM}}(\Omega) - V_{\text{eff}}$ as $C^\top \Sigma^{-1} C$ for a suitable matrix C.)

(b) Under homoskedasticity, show that 2SLS is the efficient GMM estimator by verifying Ω_{2SLS} is a scalar multiple of Σ^{-1}.

(c) Return to the two-instrument example. At $\hat{\beta} \approx 0.764$ with $\hat{\Sigma} = I_2$ and $n = 200$, compute $J = n \hat{U}(\hat{\beta})^\top \hat{\Sigma}^{-1} \hat{U}(\hat{\beta})$ and determine using the χ^2_1 critical value at the 5% level whether the overidentifying restriction is rejected.

4. GEL first-order conditions and the minimum-discrepancy dual.

(a) For EL, $G(v) = -\log(1 - v)$. Write the first-order condition for the inner supremum and show it implies $\sum_i \hat{\pi}_i U_i(\theta) = 0$ where $\hat{\pi}_i \propto (1 - \hat{\lambda}^\top U_i(\theta))^{-1}$.

(b) For CUE, $G(v) = v + v^2/2$. Solve the inner supremum explicitly at fixed θ to show $\hat{\lambda} = -[n^{-1} \sum_i U_i(\theta) U_i(\theta)^\top]^{-1} \bar{U}(\theta)$, and confirm the profile objective equals $\frac{1}{2} \bar{U}^\top [n^{-1} \sum_i U_i U_i^\top]^{-1} \bar{U}$.

(c) In the exactly identified case ($m = k$), show that $\hat{\lambda} = 0$ at any GEL solution $\hat{\theta}$. Conclude that every GEL estimator coincides with the just-identified GMM estimator and the empirical probabilities all equal $1/n$.

5. Control function, endogeneity testing, and the limits of instrument diagnostics.

Let $\varepsilon = \rho \eta + \xi$ with $\rho = \text{Cov}(\varepsilon, \eta)/\text{Var}(\eta)$.

(a) Show $\mathbb{E}[\xi] = 0$ and $\text{Cov}(\xi, \eta) = 0$ by construction. Then assume $\mathbb{E}[\varepsilon \mid \eta, Z] = \rho \eta$ and verify $\mathbb{E}[\xi \mid \eta, Z] = 0$. Explain why this renders T exogenous in the augmented regression of Y on (T, η).

(b) Show that the coefficient on $\hat{\eta}$ in the augmented regression is a consistent estimator of ρ, and connect the t-test on $\hat{\eta}$ to the Hausman (1978) endogeneity test.

(c) Suppose an instrument Z affects wages both through education and through a direct network effect, but the model is exactly identified. Explain why neither the first-stage F-test nor the control function endogeneity test can detect this exclusion-restriction violation, and what additional information would be needed.

Appendix A

Graphical Intuition for Conditional Independence and d-Separation

This appendix provides a gentle introduction to the probabilistic and graphical ideas that underlie Chapters 2 and 3. Our goal is not to give a complete treatment of graphical models, but rather to develop enough intuition so that the formal machinery of d-separation, back-door adjustment, and intervention graphs does not appear abruptly.

A directed acyclic graph (DAG) is more than a picture. It is a compact language for expressing assumptions about how variables are related. Once those assumptions are represented graphically, the graph tells us which variables may be associated, which paths transmit dependence, and which variables should or should not be conditioned on. These ideas become central in Chapter 2, where we study d-separation, and in Chapter 3, where we use graph surgery to derive identification results.

The key pedagogical idea of this appendix is simple: before learning the full d-separation criterion, it is helpful to understand three elementary three-node patterns. Those local patterns explain most of what happens later in larger graphs.

A.1 Conditional Independence: The Probabilistic Language Behind Graphs

Before introducing graphs, we first recall the probabilistic notion that graphs are designed to encode.

Definition: Conditional Independence

Let X, Y, and Z be random variables. We say that X and Y are *conditionally independent given* Z, written $X \perp\!\!\!\perp Y \mid Z$, if

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z) \qquad \text{for } P_Z\text{-almost every } z.$$

Equivalently, once Z is known, learning X gives no additional information about Y, and learning Y gives no additional information about X.

In terms of densities (with respect to a dominating measure on (X, Y, Z)), conditional independence admits the following equivalent characterizations, each interpreted almost everywhere:

$$X \perp\!\!\!\perp Y \mid Z \iff f(x, y, z)\, f(z) = f(x, z)\, f(y, z) \iff \exists\, a, b \colon f(x, y, z) = a(x, z)\, b(y, z).$$

The last form is especially useful: it says that the joint density factors into one piece depending on (x, z) and another depending on (y, z), with no cross-term in x and y.

Proposition: Fundamental Properties of Conditional Independence [@dawid1979conditional; @dawid1980conditional]

For random variables X, Y, Z, and W, the following hold:

- **(C1) Symmetry.** $X \perp\!\!\!\perp Y \mid Z \Rightarrow Y \perp\!\!\!\perp X \mid Z$.
- **(C2) Decomposition.** $X \perp\!\!\!\perp (Y, W) \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$.
- **(C3) Weak Union.** $X \perp\!\!\!\perp (Y, W) \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid (Z, W)$.
- **(C4) Contraction.** $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z) \Rightarrow X \perp\!\!\!\perp (Y, W) \mid Z$.
- **(C5) Intersection.** If $f(x, y, z, w) > 0$ for all (x, y, z, w), then $X \perp\!\!\!\perp Y \mid (Z, W)$ and $X \perp\!\!\!\perp Z \mid (Y, W) \Rightarrow X \perp\!\!\!\perp (Y, Z) \mid W$.

Properties (C1)–(C4) hold for any probability distribution and are known as the *semigraphoid axioms*. Property (C5) additionally requires the joint density to be strictly positive; together with (C1)–(C4) it forms the *graphoid axioms*. These properties are used implicitly throughout the course whenever conditional independence statements are combined or simplified.

Conditional independence is not the same as marginal independence. Two variables may be dependent marginally but independent after conditioning on a third variable. Conversely, two variables may be independent marginally but become dependent after conditioning. Both phenomena occur repeatedly in causal inference.

Example: Ice cream, drowning, and season

Let X = ice cream sales, Y = drowning incidents, and Z = season. Marginally, X and Y are positively associated because both tend to be higher in summer. This does not mean that ice cream sales cause drowning. A more plausible explanation is that season is a common cause of both variables. Once season is fixed, the association largely disappears: $X \perp\!\!\!\perp Y \mid Z$.

This example illustrates a recurring theme in causal inference: an observed association may be induced by a third variable, and conditioning on that variable can remove the spurious dependence.

Remark: The Central Question

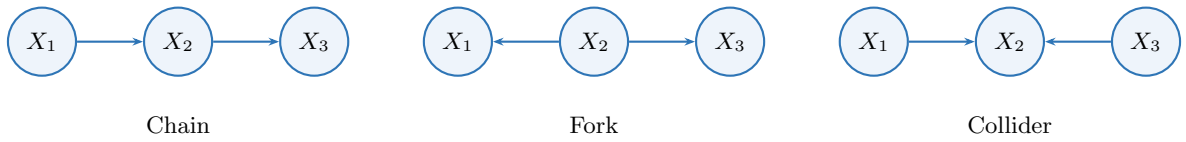
At this stage, the main question for the reader is: *Does conditioning on a variable remove association, preserve it, or create it?* Graphs provide a systematic answer to exactly this question.

A.2 Three Basic Motifs: Chain, Fork, and Collider

Every path in a DAG is built from local three-node configurations. There are three fundamental types: a *chain*, a *fork*, and a *collider*. Their behavior under conditioning is the foundation of d-separation.

```
\usetikzlibrary{arrows.meta, positioning}
\definecolor{isublue}{RGB}{46,117,182}
\definecolor{defbg}{RGB}{238,244,251}
\tikzset{node/.style={circle,draw=isublue,fill=defbg,thick,minimum size=9mm,font=\small}}
\tikzset{edge/.style={-{Stealth[length=4pt]},thick,color=isublue}}
\begin{tikzpicture}
  \node[node] (x1) at (0.0,0){$X_1$};
  \node[node] (x2) at (1.8,0){$X_2$};
  \node[node] (x3) at (3.6,0){$X_3$};
  \draw[edge](x1)--(x2); \draw[edge](x2)--(x3);
  \node[font=\small,below=4mm of x2]{Chain};
  \node[node] (y1) at (5.5,0){$X_1$};
  \node[node] (y2) at (7.3,0){$X_2$};
  \node[node] (y3) at (9.1,0){$X_3$};
  \draw[edge](y2)--(y1); \draw[edge](y2)--(y3);
  \node[font=\small,below=4mm of y2]{Fork};
  \node[node] (z1) at (11.0,0){$X_1$};
```

```
  \node[node] (z2) at (12.8,0){$X_2$};
  \node[node] (z3) at (14.6,0){$X_3$};
  \draw[edge](z1)--(z2); \draw[edge](z3)--(z2);
  \node[font=\small,below=4mm of z2]{Collider};
\end{tikzpicture}
```



Chain Fork Collider

Figure A.1: The three fundamental three-node motifs.

A.2.1 Chain

Consider the pattern $X_1 \to X_2 \to X_3$, where the middle node X_2 lies on a directed pathway from X_1 to X_3.

> Example: Exercise, body weight, and blood pressure
>
> Let X_1 = exercise, X_2 = body weight, and X_3 = blood pressure. Exercise may affect blood pressure partly through its effect on body weight. Marginally, exercise and blood pressure are associated; conditioning on body weight blocks this particular pathway. In the isolated three-node DAG, $X_1 \perp\!\!\!\perp X_3 \mid X_2$.

Once the middle variable is fixed, the chain no longer transmits additional information from X_1 to X_3.

A.2.2 Fork

Consider the pattern $X_1 \leftarrow X_2 \to X_3$, where the middle node X_2 is a common cause of X_1 and X_3.

> Example: Ice cream, season, and drowning
>
> With X_1 = ice cream sales, X_2 = season, and X_3 = drowning incidents, season affects both variables, so X_1 and X_3 are associated even though neither causes the other. Conditioning on season blocks this path: $X_1 \perp\!\!\!\perp X_3 \mid X_2$.

A fork is the simplest graphical form of confounding: the middle node creates association, and conditioning on it blocks the path.

A.2.3 Collider

Consider the pattern $X_1 \to X_2 \leftarrow X_3$, where the middle node X_2 is a common effect of X_1 and X_3.

> Example: Talent, legacy status, and college admission
>
> Let X_1 = academic talent, X_3 = legacy status, and X_2 = admission to an elite university. Talent and legacy status may be unrelated in the general applicant pool. But among admitted students they can become statistically associated: low legacy status makes unusually high talent more likely among admitted applicants, and vice versa. Symbolically, $X_1 \perp\!\!\!\perp X_3$, but conditioning on X_2 opens the path, so X_1 and X_3 are typically dependent given X_2.

Unlike chains and forks, a collider *blocks* the path by default. Conditioning on the collider opens the path and may induce association that was not present marginally.

A.2.4 Summary of the Three Motifs

A chain ($X_1 \to X_2 \to X_3$) and a fork ($X_1 \leftarrow X_2 \to X_3$) are each open by default and blocked by conditioning on the middle node X_2. A collider ($X_1 \to X_2 \leftarrow X_3$) is blocked by default and opened by conditioning on the middle node or any of its descendants.

Remark: Path-Blocking versus Conditional Independence

The independence statements in the chain, fork, and collider examples above are read in the isolated three-node DAGs as drawn. In a larger DAG, conditioning on the middle node X_2 of a chain or fork blocks *that particular path*, but X_1 and X_3 are conditionally independent only if every other path between them is also blocked.

Conditioning Is Not Always Beneficial

Many adjustment mistakes arise from forgetting this asymmetry. Conditioning on a collider can create bias rather than remove it. The informal rule "control for more variables" is unsafe in causal inference; d-separation teaches a more precise lesson: condition on the *right* variables, not simply on many variables.

Chapter 2 develops these three motifs into the full d-separation criterion; see in particular the blocking rules and the extended treatment of collider bias.

A.3 d-Separation: When Does Conditioning Block a Path?

The three-node motifs explain what happens locally on a path. The next step is to extend this logic to a general DAG, where two variables may be connected by many paths.

Definition: Path

A *path* between two nodes is a sequence of distinct nodes such that each consecutive pair is connected by an edge, regardless of edge direction.

Definition: Blocked Path

A path is *blocked* by a conditioning set S if at least one of the following holds: (1) the path contains a chain or fork node that belongs to S, or (2) the path contains a collider such that neither the collider nor any of its descendants belongs to S. A path is *open* given S if it is not blocked.

Definition: d-Separation

Two nodes X and Y are *d-separated* by a set S if every path between X and Y is blocked by S. More generally, two disjoint sets of nodes $\mathbf{X}$ and $\mathbf{Y}$, each disjoint from S, are d-separated by S if every $X \in \mathbf{X}$ is d-separated from every $Y \in \mathbf{Y}$ by S.

Remark: d-Separation versus Conditional Independence

d-Separation is a graphical condition; conditional independence is a probabilistic one. The Markov property guarantees only one direction: d-separation in $\mathcal{G}$ implies the corresponding conditional independence in every distribution that factorizes according to $\mathcal{G}$. The reverse implication — that d-connection forces conditional dependence — requires a faithfulness or no-cancellation assumption. We therefore read an open path as "the graph does not force independence," not as "dependence is guaranteed." This distinction is taken up in detail in Chapter 2.

A.3.1 A Confounding Example

Consider the DAG with edges $X \to T$, $T \to Y$, and $X \to Y$. Here T is the treatment, Y is the outcome, and X is a pre-treatment covariate that affects both. There are two paths from T to Y: the directed causal path $T \to Y$, and the back-door path $T \leftarrow X \to Y$.

Without conditioning, the back-door path is open, so the observed association between T and Y mixes the causal effect with confounding. Conditioning on X blocks the fork $T \leftarrow X \to Y$, thereby isolating the causal path. This confounding graph anticipates the back-door criterion of Chapter 3: the graphical condition that makes adjustment valid is precisely that all back-door paths are blocked.

A.3.2 A Collider Warning

Now consider the DAG with edges $T \to C$, $U \to C$, and $U \to Y$. Here C is a collider on the path $T \to C \leftarrow U \to Y$. Without conditioning on C, the path is blocked at the collider. Conditioning on C opens the path, creating a spurious association between T and Y through U.

Even more subtly, conditioning on a *descendant* of C can also open the path: if additionally $C \to D$, then conditioning on D may also induce association between T and Y through the collider at C.

A.3.3 A Practical Checklist

To decide whether X and Y are d-separated by S, proceed as follows. First, list all paths between X and Y. On each path, classify each interior node as part of a chain, fork, or collider. Then check whether each path is blocked by S. Finally, conclude that X and Y are d-separated if and only if every path is blocked.

A.4 DAG Factorization and the Markov Property

Up to this point, we have used DAGs qualitatively, to decide which paths are open or blocked. We now connect the graph to probability algebra.

Definition: DAG Factorization

Let $\mathcal{G}$ be a DAG with nodes $V_1, \dots, V_p$. A joint distribution $p(v_1, \dots, v_p)$ is said to *factorize according to* $\mathcal{G}$ if

$$p(v_1, \dots, v_p) = \prod_{j=1}^{p} p\big(v_j \mid \mathrm{Pa}(V_j)\big)\,,$$

where $\mathrm{Pa}(V_j)$ denotes the set of parents of V_j in $\mathcal{G}$.

This factorization implies the local Markov property: once its parents are known, a node is conditionally independent of all variables that are neither its descendants nor its parents.

Example: A simple confounding graph

For the DAG with edges $X \to T$, $T \to Y$, and $X \to Y$, the joint density factorizes as $p(x, t, y) = p(x)\, p(t \mid x)\, p(y \mid t, x)$.

Definition: Local Markov Property

Let $\mathrm{Nd}(V_i) = V \setminus \big(\{V_i\} \cup \mathrm{De}(V_i)\big)$ denote the set of *non-descendants* of V_i. A distribution P satisfies the *local Markov property* with respect to $\mathcal{G}$ if, for every node $V_i \in V$,

$$V_i \perp\!\!\!\perp \big(\mathrm{Nd}(V_i) \setminus \mathrm{Pa}(V_i)\big) \mid \mathrm{Pa}(V_i).$$

Remark: Global Markov Property

The global Markov property is the statement that every d-separation in $\mathcal{G}$ implies a conditional independence in P. This is studied in detail in Chapter 2.

Remark: Equivalence for DAGs

For DAGs, under the existence of regular conditional distributions, the recursive factorization, the local Markov property, and the global Markov property are all equivalent. The local property is the version most commonly verified in practice, since the factorization makes it immediate.

Example: Education–earnings graph

Consider the DAG with edges $N \to E$, $B \to E$, $E \to Y$, and $B \to Y$ (N = neighborhood, B = family background, E = education, Y = earnings). The corresponding factorization is $p(n, b, e, y) = p(n)\,p(b)\,p(e \mid n, b)\,p(y \mid e, b)$. The parent set of Y is $\mathrm{Pa}(Y) = \{E, B\}$, so the local Markov property gives $Y \perp\!\!\!\perp N \mid \{E, B\}$. The factorization also implies $N \perp\!\!\!\perp B$, since neither node has a common ancestor; this is a substantive modeling assumption.

Remark: Edges as Scientific Claims

In a causal DAG, an arrow is typically drawn only when a direct dependence-generating relation is believed to be present. Every arrow represents a substantive scientific claim. Minimality and faithfulness are discussed in Chapter 2.

Optional: Moralization as an Alternative Criterion

Note to Reader

This section may be skipped on a first reading.

There is an alternative graph-theoretic way to check d-separation based on constructing an undirected graph called the *moral graph*. To check whether $X \perp\!\!\!\perp Y \mid S$: First, take the induced subgraph on $\mathrm{An}(X \cup Y \cup S)$, the ancestors of all variables under consideration. Second, connect any two parents of a common child by an undirected edge. Third, drop all arrow directions. Finally, check whether S separates X and Y in the resulting undirected graph. This procedure yields a criterion equivalent to d-separation.

Example: Moral graph of a collider

Consider the DAG $A \to C \leftarrow B$. For the conditional query $A \perp\!\!\!\perp B \mid C$, the ancestral set is $\{A, B, C\}$. Moralization connects the two parents A and B, and after deleting the conditioned node C the edge $A - B$ remains; therefore A and B are not separated, matching the fact that conditioning on a collider opens the path. By contrast, for the marginal query $A \perp\!\!\!\perp B$, the ancestral set is $\{A, B\}$, so C is discarded before moralization and no moral edge is added; A and B are separated, as expected.

Summary

This appendix introduced the graphical ideas that support Chapters 2 and 3. Conditional independence is the probabilistic language that graphs are designed to encode. The three local motifs — chain, fork, and collider — determine how conditioning affects association along a path. d-Separation extends these local rules to arbitrary graphs: two variables are d-separated by S if every path between them is blocked by S. The most important application in causal inference is to distinguish confounding paths from causal paths and to identify valid adjustment sets. Finally, the Markov property gives a probabilistic interpretation to the graph by linking graphical structure to a factorization of the joint distribution.

Appendix B

Single World Intervention Graphs

This appendix develops single world intervention graphs (SWIGs), a graphical device introduced by Richardson and Robins (2014) and given an accessible practical treatment by Bezuidenhout et al. (2025). A SWIG provides a formal bridge between the potential outcomes framework and the do-calculus by encoding both in a single diagram. A reader familiar with Chapters 2–4 will find that SWIGs require no new conceptual ingredients: they are ordinary DAGs in which the treatment node is split into two pieces, making potential outcomes and identifiability assumptions graphically explicit rather than leaving them implicit.

Chapter 4 used the back-door criterion and the adjustment formula to identify causal effects. SWIGs make that identification argument graphically explicit: potential outcomes appear as labeled nodes inside the graph, so ignorability becomes a routine d-separation statement rather than a free-standing probabilistic assumption. Section B.5 restates the back-door identification result of Chapter 4 as a graph-theoretic theorem inside the SWIG, and examines a partial converse: under faithfulness of the SWIG and a restriction of the adjustment set to nondescendants of the treatment, the single-world ignorability statement $Y(t) \perp\!\!\!\perp T \mid X$ entails the back-door criterion.

B.1 The SWIG Construction

In the original DAG $\mathcal{G}$, the node T represents the treatment as it naturally occurs. Under do(T=t), however, the treatment is externally fixed. The SWIG separates these two roles by *splitting* T into two pieces displayed side by side in the graph:

- a **random half** (left side, capital letter T): represents the treatment value that would naturally occur, and retains all incoming edges from T's causal parents;
- a **fixed half** (right side, lowercase letter t): represents the externally imposed intervention value, carries all outgoing edges that formerly left T, and has *no* edge to or from the random half.

Definition: Single World Intervention Graph [@richardson2014ace]

The *SWIG* $\mathcal{G}(t)$ is constructed from $\mathcal{G}$ by:

1. Replacing T with two nodes: random half T (all incoming edges retained) and fixed half t (all outgoing edges retained).
2. Severing the direct connection between the two halves.
3. Redrawing causal arrows at the split node: *all arrows into the split node land on the random half*; *all arrows departing the split node leave from the fixed half.*
4. Relabeling every descendant V of T in $\mathcal{G}$ as $V(t)$, indicating it is a potential outcome under do(T=t).

```
\usetikzlibrary{arrows.meta,positioning}
\definecolor{isublue}{RGB}{46,117,182}
\definecolor{defbg}{RGB}{238,244,251}
\tikzset{rnd/.style={circle,draw=isublue,fill=defbg,thick,minimum size=9mm,font=\small\bfseries},
```

```
        fix/.style={rectangle,draw=isublue,fill=gray!15,thick,minimum size=9mm,font=\small\bfseries}
\begin{tikzpicture}[>=stealth]
  \begin{scope}
    \node[rnd](L) at (1.4,1.5){$L$};
    \node[rnd](A) at (0,0){$A$};
    \node[rnd](Y) at (2.8,0){$Y$};
    \draw[->,thick,color=isublue](L)--(A); \draw[->,thick,color=isublue](L)--(Y); \draw[->,thick,colo
    \node[font=\small\itshape,below=5mm of A,xshift=14mm]{DAG};
  \end{scope}
  \node[font=\large] at (3.9,0.4){$\Rightarrow$};
  \begin{scope}[xshift=4.9cm]
    \node[rnd](Lr) at (1.9,1.5){$L$};
    \node[rnd](Ar) at (0.4,0){$A$};
    \node[fix](af) at (1.5,0){$a$};
    \node[rnd](Ya) at (3.5,0){$Y(a)$};
    \draw[->,thick,color=isublue](Lr)--(Ar); \draw[->,thick,color=isublue](Lr)--(Ya); \draw[->,thick
    \node[font=\small\itshape,below=5mm of Ar,xshift=16mm]{SWIG $\mathcal{G}(a)$};
  \end{scope}
\end{tikzpicture}
```

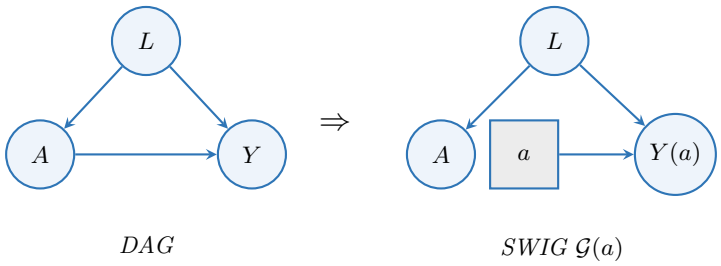


Figure B.1: Converting the confounded-treatment DAG into the SWIG for do(A=a).

B.2 The Fixed Half as a Source Node

After the SWIG $\mathcal{G}(t)$ is constructed, d-separation among its nodes is evaluated using the standard rules of Chapter 2. The key structural fact is:

The Fixed Half Is a Source Node

In $\mathcal{G}(t)$, **no directed edge points into** the fixed half t. Two structural consequences follow.

1. **The route $T \to t \to Y(t)$ does not exist.** Because no edge connects the random half T to the fixed half t, the directed concatenation $T \to t \to Y(t)$ is *not a path in the graph* — there is nothing for d-separation to block.
2. **Every path from T to a potential outcome $V(t)$ leaves T through one of T's natural causes.** The only edges incident to the random half T in $\mathcal{G}(t)$ are the incoming edges from T's parents in $\mathcal{G}$. Any path from T to $V(t)$ must therefore start with an arrow pointing *into* T — that is, every such path is a back-door path.

"Source Node" Means No Parents, Not "On No Path"

Calling t a source node refers to the absence of *directed edges into* t, nothing more. An undirected path can still pass through a source node. What the source-node property delivers is the asymmetry that matters for the back-door argument: at the random half T, *all* edges point inward, so any path leaving T must do so through one of T's causes.

Reading the back-door criterion off the SWIG. By the source-node analysis above, every path from the random half T to a potential outcome $V(t)$ in $\mathcal{G}(t)$ is a back-door path in $\mathcal{G}$, so blocking all back-door

paths is exactly what d-separation of T and $V(t)$ in the SWIG asks for.

Recovering ignorability from the SWIG. In the SWIG of the confounded-treatment example, the only path from the random half A to $Y(a)$ is $A \leftarrow L \rightarrow Y(a)$: open, but blocked by conditioning on L. Conditioning on L achieves d-separation, and the standard rules give $A \perp\!\!\!\perp Y(a) \mid L$ in $\mathcal{G}(a)$, which is exactly the conditional ignorability of Chapter 4, now *derived* as a graph-structural consequence rather than asserted as an assumption.

B.3 When Ignorability Fails: Hidden Confounding

The SWIG is equally useful for diagnosing *failures* of ignorability. Suppose the DAG is augmented by a hidden common cause U that affects both A and Y but is not recorded in the data.

```
\usetikzlibrary{arrows.meta,positioning}
\definecolor{isublue}{RGB}{46,117,182}
\definecolor{defbg}{RGB}{238,244,251}
\tikzset{rnd/.style={circle,draw=isublue,fill=defbg,thick,minimum size=9mm,font=\small\bfseries},
         fix/.style={rectangle,draw=isublue,fill=gray!15,thick,minimum size=9mm,font=\small\bfseries
         hid/.style={circle,draw=isublue,fill=white,thick,dashed,minimum size=9mm,font=\small\bfseri
\begin{tikzpicture}[>=stealth]
  \begin{scope}
    \node[rnd](L) at (1.4,1.5){$L$}; \node[rnd](A) at (0,0){$A$}; \node[rnd](Y) at (2.8,0){$Y$}; \no
    \draw[->,thick,color=isublue](L)--(A); \draw[->,thick,color=isublue](L)--(Y); \draw[->,thick,col
    \draw[->,thick,dashed,color=isublue](U)--(A); \draw[->,thick,dashed,color=isublue](U)--(Y);
    \node[font=\small\itshape,below=3mm of U,xshift=0]{DAG (with $U$)};
  \end{scope}
  \node[font=\large] at (3.9,0.15){$\Rightarrow$};
  \begin{scope}[xshift=4.9cm]
    \node[rnd](Lr) at (1.9,1.5){$L$}; \node[rnd](Ar) at (0.4,0){$A$}; \node[fix](af) at (1.5,0){$a$}
    \draw[->,thick,color=isublue](Lr)--(Ar); \draw[->,thick,color=isublue](Lr)--(Ya); \draw[->,thick
    \draw[->,thick,dashed,color=isublue](Ur)--(Ar); \draw[->,thick,dashed,color=isublue](Ur)--(Ya);
    \node[font=\small\itshape,below=3mm of Ur,xshift=0]{SWIG (with $U$)};
  \end{scope}
\end{tikzpicture}
```

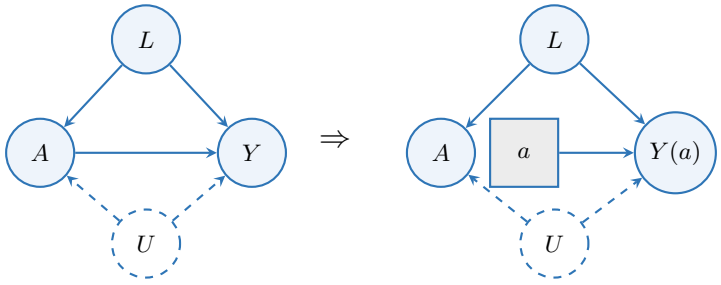


Figure B.2: Adding a hidden confounder U (dashed circle). Ignorability fails.

In the SWIG $\mathcal{G}(a)$ with hidden U, the paths from A to $Y(a)$ are: (1) $A \leftarrow L \rightarrow Y(a)$: blocked by conditioning on L; and (2) $A \leftarrow U \rightarrow Y(a)$: open regardless of L, since U is unobserved and cannot be conditioned on. Even after conditioning on L, $A \not\perp\!\!\!\perp Y(a) \mid L$, so ignorability fails and no adjustment for observed covariates alone can identify the ATE. The failure is *immediately visible* from the SWIG: an unblocked dashed arrow entering the random half A signals that selection into treatment carries information about the potential outcome that no set of observed covariates can remove.

Remark

This failure motivates the instrumental variable designs of Chapter 7. When a valid instrument Z is available — one that affects A but has no direct effect on Y and is independent of U — the IV strategy sidesteps the unblocked U-path rather than attempting to block it by covariate adjustment.

B.4 What SWIGs Achieve

1. Potential outcomes as random variables in a DAG. In $\mathcal{G}(t)$, the variable $Y(t)$ is an ordinary node with well-defined parents: the fixed half t and any other causes of Y retained from $\mathcal{G}$. Its marginal distribution equals the interventional distribution $f(y \mid \text{do}(T{=}t))$.

2. The mutilated graph as a special case. Removing the random half T from $\mathcal{G}(t)$, and suppressing the fixed-half labeling on its descendants, recovers the usual mutilated graph $\mathcal{G}_{\overline{T}}$ of Chapter 1 representing $\text{do}(T{=}t)$. The SWIG is more informative: it retains the random half alongside the fixed half, making it possible to ask questions about the relationship between the natural treatment T and the potential outcomes $V(t)$ via d-separation.

3. Unconfoundedness as a d-separation statement. The ignorability condition $Y(t) \perp\!\!\!\perp T \mid X$ corresponds, under the Markov property of the SWIG, to a d-separation statement *in* $\mathcal{G}(t)$ verifiable by path-tracing. Ignorability is no longer an assertion encoded in potential-outcome notation alone; it is a structural claim about the graph.

4. A single graph for both frameworks. Before SWIGs, combining the potential outcomes framework and the do-calculus required switching notation. The SWIG eliminates this translation step: one graph simultaneously hosts the natural treatment T (random half), the intervention value t (fixed half), the potential outcomes $V(t)$, and all d-separation relationships among them.

Cross-World Independence and Its Limits

Consider $Y(1) \perp\!\!\!\perp Y(0)$: $Y(1)$ lives in SWIG $\mathcal{G}(1)$ and $Y(0)$ lives in SWIG $\mathcal{G}(0)$. No unit is ever observed in both worlds simultaneously, so no data can directly test this independence. A more substantive cross-world example arises in mediation analysis (Chapter 8): identification of natural direct and indirect effects relies on assumptions of the form $Y(t, m) \perp\!\!\!\perp M(t') \mid X$ $(t \neq t')$, which combine potential outcomes from *different SWIGs*. SWIGs make such assumptions explicit: the analyst must specify a joint distribution across multiple SWIGs. The single-world results of Section B.5 therefore cannot themselves supply cross-world assumptions.

Remark

More than one node may be split in a single SWIG. In longitudinal settings with time-varying treatments (A_0, A_1), splitting both nodes yields the sequential-randomization SWIG, from which the two *sequential exchangeability* conditions needed to identify the g-formula follow directly by d-separation.

B.5 Single-World Ignorability and the Back-Door Criterion

This section restates the back-door identification result of Chapter 4 graphically as a theorem inside the SWIG, and then examines a partial converse. The target conditional independence is the *single-world* (or *weak*) ignorability statement $Y(t) \perp\!\!\!\perp T \mid X$ $(t = 0, 1)$. The cross-world joint statement $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$ involves $Y(0)$ and $Y(1)$, which live in two distinct SWIGs; a single SWIG cannot encode it as a d-separation.

Definition: Faithfulness of the SWIG

Let $\mathcal{G}(t)$ be the SWIG induced by an intervention do(T=t), and let P_t denote the joint distribution it induces over the random nodes. The distribution P_t is **faithful** to $\mathcal{G}(t)$ if every conditional independence that holds in P_t corresponds to a d-separation in $\mathcal{G}(t)$:

$$A \perp\!\!\!\perp B \mid C \text{ in } P_t \implies A \text{ and } B \text{ are d-separated by } C \text{ in } \mathcal{G}(t).$$

Faithfulness rules out accidental cancellations of causal paths in the structural model. In smooth parametric families with a fixed graph, the parameter configurations that produce unfaithfulness typically form a lower-dimensional subset, so faithfulness is, in that sense, generic. The qualifier matters: deterministic relationships, context-specific independencies, and graph misspecification can all produce substantive failures of faithfulness.

Faithfulness is invoked only for the converse direction below.

Theorem: Back-Door Criterion Implies Single-World Ignorability

Fix an intervention value t. Suppose the structural causal model underlying $\mathcal{G}$ induces a distribution P_t over the random nodes of the SWIG $\mathcal{G}(t)$ that is Markov with respect to $\mathcal{G}(t)$. If X contains no descendants of T in $\mathcal{G}$ and blocks every back-door path from T to Y, then $Y(t) \perp\!\!\!\perp T \mid X$.

Proof. By the source-node analysis of Section B.2, every path from the random half T to the potential outcome $Y(t)$ in $\mathcal{G}(t)$ leaves T through one of T's parents in $\mathcal{G}$ and is therefore a back-door path. By hypothesis, X blocks every such path; the no-descendants clause ensures that conditioning on members of X does not open a previously blocked collider path. Hence T and $Y(t)$ are d-separated by X in $\mathcal{G}(t)$, and the Markov property of P_t delivers $Y(t) \perp\!\!\!\perp T \mid X$. □

For binary treatment, applying the theorem separately at $t = 0$ and $t = 1$ yields the weak ignorability conditions $Y(0) \perp\!\!\!\perp T \mid X$ and $Y(1) \perp\!\!\!\perp T \mid X$ that justify the adjustment formula of Chapter 4.

Remark: A Partial Converse

If X is restricted *a priori* to pre-treatment (nondescendant) variables in $\mathcal{G}$ and P_t is faithful to $\mathcal{G}(t)$, the implication above can be partially reversed:

$$Y(t) \perp\!\!\!\perp T \mid X \text{ in } P_t \implies X \text{ blocks every back-door path from } T \text{ to } Y.$$

By faithfulness, the assumed independence corresponds to d-separation of the random half T and $Y(t)$ by X in $\mathcal{G}(t)$; since every path from T to $Y(t)$ is a back-door path, X must block all of them. Without the nondescendant restriction, however, conditional exchangeability alone does not imply the full back-door criterion.

Why "No Descendants of T" Is a Hypothesis, Not a Conclusion

The partial converse imposes "no descendants of T" as a hypothesis. This is essential: the conditional independence $Y(t) \perp\!\!\!\perp T \mid X$ does *not* by itself imply that X contains no descendant of T. For instance, take the DAG with edges $T \to Y$ and $T \to D$ and no confounding, and let $X = \{D\}$. D is a descendant of T, so X violates the back-door criterion. Nonetheless, $Y(t) \perp\!\!\!\perp T \mid D$ may continue to hold because there is no back-door path from T to Y at all. Conditioning on a descendant of T *may* open a previously blocked collider path, but it need not. The no-descendants clause belongs to the *definition* of an admissible graphical adjustment set and cannot be inferred from observed conditional independencies alone.

From single-world to cross-world. The strong ignorability condition $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$ is a statement about the joint distribution of $Y(0)$ and $Y(1)$ together with T. Because $Y(0)$ and $Y(1)$ are nodes in two distinct SWIGs, no single SWIG represents this joint distribution as a d-separation, and the theorem above correspondingly speaks only about the single-world form. For ATE identification this is no obstacle: the single-world form $Y(t) \perp\!\!\!\perp T \mid X$ for each $t \in \{0, 1\}$ is enough. Identification of joint or

distributional functionals of $(Y(0), Y(1))$ — for instance, the variance of the unit-level treatment effect, or the proportion of units harmed by treatment — requires the cross-world joint form and assumptions across multiple SWIGs that single-world graphical reasoning cannot supply.

Conceptual takeaway. The back-door criterion is a graphical sufficient condition for the single-world conditional exchangeability assumption needed for ATE identification. Under the additional restriction that the candidate adjustment set is composed of nondescendants of T, and under faithfulness of the relevant counterfactual distribution to the SWIG, the absence of unblocked back-door paths corresponds to the single-world independence $Y(t) \perp\!\!\!\perp T \mid X$. In this sense the graphical and counterfactual languages describe the same identifying structure. Strict logical equivalence in full generality is more delicate: ignorability can hold in particular models for parametric reasons that no graphical criterion will detect.

Problems

1. SWIG construction. Consider the DAG: $X \to T$, $X \to Y$, $T \to Y$, with X fully observed.

(a) Construct the SWIG $\mathcal{G}(t)$ by splitting T into its random and fixed halves. Draw the result, labeling the random half, fixed half, and the potential outcome $Y(t)$.
(b) In $\mathcal{G}(t)$, identify all paths between the random half T and $Y(t)$. Apply d-separation to determine which are blocked and which are open before conditioning.
(c) Verify that X satisfies the back-door criterion in $\mathcal{G}$: show that X blocks the unique back-door path $T \leftarrow X \to Y$ and that X contains no descendant of T. Apply the theorem above to conclude $Y(t) \perp\!\!\!\perp T \mid X$, and confirm that the same conclusion can be read directly off $\mathcal{G}(t)$ via d-separation.
(d) Now add a hidden common cause $U \to T$, $U \to Y$. Draw the revised SWIG. Does the ignorability argument still hold? State the correct conclusion and identify what additional structure (if any) would be needed for identification.

2. Sequential exchangeability in a longitudinal SWIG. Suppose the treatment is time-varying: T_0 is assigned at baseline and T_1 at a second time point, and a time-varying covariate L is measured between the two assignments. The DAG has edges $X \to T_0$, $X \to L$, $X \to Y$, $T_0 \to L$, $T_0 \to T_1$, $T_0 \to Y$, $L \to T_1$, $L \to Y$, and $T_1 \to Y$.

(a) Construct the SWIG $\mathcal{G}(t_0, t_1)$ by splitting both T_0 and T_1. Label the random halves, the fixed halves t_0 and t_1, and the potential outcomes. In particular, the random half of T_1 should be labeled $T_1(t_0)$ and the time-varying covariate should be labeled $L(t_0)$.
(b) List the back-door paths in $\mathcal{G}(t_0, t_1)$ from the random half T_0 to $Y(t_0, t_1)$, and from $T_1(t_0)$ to $Y(t_0, t_1)$. For each, identify the smallest subset of $\{X, T_0, L(t_0)\}$ whose conditioning achieves d-separation.
(c) State the two sequential exchangeability conditions that the SWIG makes graphically transparent: $Y(t_0, t_1) \perp\!\!\!\perp T_0 \mid X$ and $Y(t_0, t_1) \perp\!\!\!\perp T_1(t_0) \mid X, T_0 = t_0, L(t_0)$. Explain in plain language what each says about treatment assignment at the corresponding time point.
(d) In observed-data shorthand, the second condition is often written $Y(\bar{t}) \perp\!\!\!\perp T_1 \mid X, T_0 = t_0, L$. State the implicit consistency step that licenses replacing $L(t_0)$ with the observed L on the event $\{T_0 = t_0\}$.
(e) What goes wrong if the analyst omits L from the conditioning set at the second stage? What goes wrong if the analyst conditions on L but treats it as a baseline covariate? Connect your answers to the role of $L(t_0)$ as a treatment-induced confounder.

3. Why the no-descendants clause is a hypothesis. Consider the DAG with edges $T \to Y$ and $T \to D$, no confounders, and let $X = \{D\}$.

(a) Construct the SWIG $\mathcal{G}(t)$. Identify all paths between the random half T and $Y(t)$, and determine whether any back-door paths exist.
(b) Argue that, in this graph, $Y(t) \perp\!\!\!\perp T \mid D$ holds in P_t even though $X = \{D\}$ violates the back-door criterion.
(c) Explain in your own words why this example shows that the no-descendants clause must be imposed as a hypothesis on the adjustment set rather than derived from a conditional independence statement.

Appendix C

Pathwise Differentiability and Efficient Influence Functions

Section Section 10.9 introduced the efficient influence function of the average treatment effect functional and stated, without proof, that this object is the *canonical gradient* of Ψ relative to the nonparametric tangent space. This appendix supplies the geometric machinery behind that statement. The treatment follows Bickel et al. (1993) and Tsiatis (2006), with notation aligned to Chapter 10.

The reader should think of this appendix as the formal counterpart of Sections on asymptotic linearity and efficiency in Chapter 10: those sections showed how an influence function determines the asymptotic distribution of an estimator; here we develop the parallel notion of an influence function as a *derivative of a functional*, and characterize the unique influence function that achieves the semiparametric efficiency bound. No new notation is needed beyond what is standard in modern semiparametric theory.

The applied consequences — the AIPW estimator, double robustness, Neyman orthogonality, cross-fitting, and rate conditions — are developed in Chapters 11 and 12. This appendix supplies only the foundational geometry on which those chapters rest. Section Section C.1 collects the Hilbert-space background used throughout. Section Section C.7 closes the appendix by carrying out the EIF derivation explicitly for the ATE, recovering the AIPW influence function from the projection construction of the Canonical Gradient Theorem.

C.1 Hilbert-Space Background

The geometry of pathwise differentiability lives in the Hilbert space $L_2(P)$ of square-integrable real-valued functions on $\mathcal{O}$, equipped with the inner product $\langle f, g\rangle_P = \mathbb{E}_P[f(O)\, g(O)]$ and norm $\|f\|_P = \langle f, f\rangle_P^{1/2}$. Two functions are **orthogonal**, written $f \perp g$, when $\langle f, g\rangle_P = 0$. All scores and influence functions introduced below live in the closed subspace of mean-zero functions, $L_2^0(P) = \{f \in L_2(P) : \mathbb{E}_P[f(O)] = 0\}$.

This section records the four Hilbert-space facts used repeatedly in the rest of the appendix. For a full treatment, see Vaart (1998, sec. 25.7) or the appendices of Tsiatis (2006).

Closed linear span. For a subset $A \subset L_2^0(P)$, the **closed linear span** $\overline{\text{span}}(A)$ is the smallest closed subspace of $L_2^0(P)$ containing A; it consists of $L_2(P)$-limits of finite linear combinations of elements of A. Tangent spaces are defined as closed linear spans because the set of scores of regular parametric submodels need not itself be closed.

Projection theorem. If $V \subset L_2^0(P)$ is a closed subspace, every $f \in L_2^0(P)$ admits a unique decomposition $f = f_V + f_{V^\perp}$ with $f_V \in V$ and $f_{V^\perp} \in V^\perp$, where $V^\perp = \{g \in L_2^0(P) : \langle g, h\rangle_P = 0 \text{ for all } h \in V\}$. The map $\Pi[\,\cdot \mid V\,] : f \mapsto f_V$ is the **orthogonal projection** onto V, characterized by (i) $f - \Pi[f \mid V] \perp V$ (orthogonality of the residual), or (ii) $\Pi[f \mid V]$ is the unique element of V minimizing $\|f - g\|_P$ over $g \in V$. The decomposition $L_2^0(P) = V \oplus V^\perp$ is the engine behind the canonical-gradient construction.

Riesz representation. Every continuous linear functional $\Lambda : L_2^0(P) \to \mathbb{R}$ admits a unique representer $r_\Lambda \in L_2^0(P)$ such that $\Lambda(f) = \langle r_\Lambda, f\rangle_P$ for all $f \in L_2^0(P)$. Pathwise differentiability of a functional Ψ is

precisely the statement that the score-to-derivative map $S \mapsto \partial_\varepsilon \Psi(P_\varepsilon)|_0$ extends to a continuous linear functional on the tangent space, and the influence function is its Riesz representer.

Codimension and uniqueness. A non-zero continuous linear functional Λ on a Hilbert space H has closed kernel $\ker(\Lambda) := \{f : \Lambda(f) = 0\}$ of codimension one, with $\ker(\Lambda)^\perp = \operatorname{span}\{r_\Lambda\}$ spanned by the Riesz representer. This codimension-one fact is what makes the efficient influence function of a real-valued functional unique, once it exists.

C.2 Regular Parametric Submodels and Scores

Let $\mathcal{P}$ be a statistical model for the distribution of the observed data O on $\mathcal{O}$, with true distribution $P \in \mathcal{P}$. All densities are taken with respect to a common dominating measure μ. The parameter of interest is a smooth real-valued functional $\Psi : \mathcal{P} \to \mathbb{R}$, $\psi = \Psi(P)$.

For vector-valued $\Psi : \mathcal{P} \to \mathbb{R}^p$, each component is pathwise differentiable in the sense of the definition below and has its own gradient; stacking gives a vector-valued influence function $\varphi^* = (\varphi_1^*, \dots, \varphi_p^*)^\top$ with each $\varphi_j^* \in L_2^0(P)$. The semiparametric efficiency bound is then the covariance matrix $\mathbb{E}_P[\varphi^*(O)\,\varphi^*(O)^\top] \in \mathbb{R}^{p\times p}$, and asymptotic-variance comparisons among regular asymptotically linear estimators are made in the Loewner partial order. Beyond this matricial bookkeeping no new ideas arise, and for clarity we restrict attention to $p = 1$ throughout.

Definition: Regular Parametric Submodel

A **regular parametric submodel** of $\mathcal{P}$ passing through P is a one-parameter family $\{P_\varepsilon : \varepsilon \in (-\delta, \delta)\} \subset \mathcal{P}$, $P_0 = P$, with densities p_ε such that $\varepsilon \mapsto \log p_\varepsilon(O)$ is differentiable at $\varepsilon = 0$ in $L_2(P)$, with **score**

$$S(O) = \frac{\partial}{\partial \varepsilon} \log p_\varepsilon(O)\Big|_{\varepsilon=0} \in L_2(P).$$

This L_2-differentiability condition is a convenient shorthand for the regularity assumptions — typically formalized by differentiability in quadratic mean of $\varepsilon \mapsto p_\varepsilon^{1/2}$ — under which scores are valid $L_2(P)$ directional derivatives and derivative-under-the-integral calculations are justified. It is weaker than pointwise-smoothness; see Vaart (1998, sec. 7.2) for the rigorous formulation.

A regular submodel is a smooth one-dimensional curve through $\mathcal{P}$ that can be probed by ordinary parametric methods. Because $\int p_\varepsilon \, d\mu = 1$ for all ε, differentiating under the integral gives $\mathbb{E}_P[S(O)] = 0$ and $\mathbb{E}_P[S(O)^2] < \infty$, so every score lies in $L_2^0(P)$.

Remark: Why Submodels Rather Than Point-Mass Contamination

A heuristic sometimes used for "probing" a functional is the contamination path $P_\varepsilon = (1-\varepsilon)P + \varepsilon\delta_o$. The corresponding Hampel influence function is the derivative $\partial_\varepsilon \Psi\{(1-\varepsilon)P + \varepsilon\delta_o\}|_{\varepsilon=0}$. Contamination is intuitive but does not yield a regular submodel in the sense of the definition above: when P is continuous, δ_o is not absolutely continuous with respect to P, so the Radon–Nikodym derivative $d\delta_o/dP$ does not exist as an $L_2(P)$ function, and the contamination path lacks a well-defined score in the sense used here. All formal results below use the regular-submodel definition, while the contamination heuristic remains useful for guessing the form of an influence function in concrete examples.

C.3 Pathwise Differentiability

The functional Ψ is differentiable along a submodel if the map $\varepsilon \mapsto \Psi(P_\varepsilon)$ is differentiable at $\varepsilon = 0$ in the ordinary sense. Pathwise differentiability asserts that this derivative can be represented as an inner product between a fixed function and the score, uniformly over submodels.

Definition: Pathwise Differentiability and Influence Function

The functional $\Psi : \mathcal{P} \to \mathbb{R}$ is **pathwise differentiable** at P if there exists a function $\varphi \in L_2^0(P)$ such that, for every regular parametric submodel $\{P_\varepsilon\}$ with score S,

$$\frac{\partial}{\partial \varepsilon} \Psi(P_\varepsilon)\Big|_{\varepsilon=0} = \mathbb{E}_P[\varphi(O)\, S(O)]. \tag{C.1}$$

Any such φ is called an **influence function** (or **gradient**) of Ψ at P. The set of influence functions is denoted $\mathrm{IF}(\Psi, P)$.

Equation Equation C.1 is the defining identity of semiparametric theory. It says the perturbation of Ψ along a submodel is fully encoded by the $L_2(P)$ inner product of a fixed function φ with the score S. Geometrically, φ is a representer for the linear functional $S \mapsto \partial_\varepsilon \Psi(P_\varepsilon)|_0$ restricted to the space of scores.

Example: Mean Functional

Let $O = Y$ with $\mathbb{E}_P[Y^2] < \infty$ and $\Psi(P) = \mathbb{E}_P[Y]$. For any regular submodel with score S:

$$\frac{\partial}{\partial \varepsilon} \int y\, p_\varepsilon(y)\, d\mu(y)\Big|_0 = \int y\, S(y)\, p(y)\, d\mu(y) = \mathbb{E}_P[Y \cdot S(O)].$$

Subtracting $\mathbb{E}_P[Y] \cdot \mathbb{E}_P[S] = 0$ gives $\partial_\varepsilon \Psi(P_\varepsilon)|_0 = \mathbb{E}_P[(Y - \mathbb{E}_P Y)\, S(O)]$, so $\varphi(O) = Y - \Psi(P)$ is an influence function of the mean functional. Under the nonparametric model, this is in fact the *unique* influence function, and hence the efficient influence function.

Remark: Connection to Estimation

If an estimator $\hat{\psi}$ is asymptotically linear at P with influence function φ in the sense of Chapter 10, and if $\hat{\psi}$ is regular along every submodel, then φ satisfies Equation C.1 and is therefore a gradient of Ψ. Pathwise differentiability is thus a necessary condition for the existence of a regular asymptotically linear estimator of Ψ (Tsiatis 2006, Theorem 3.1).

C.4 Non-Uniqueness of Influence Functions

Definition of pathwise differentiability does not produce a unique influence function. If φ satisfies Equation C.1 and $h \in L_2^0(P)$ is orthogonal to every score of the model in $L_2(P)$, then $\varphi + h$ also satisfies Equation C.1, since $\mathbb{E}_P[(\varphi + h)S] = \mathbb{E}_P[\varphi S] + \mathbb{E}_P[hS] = \mathbb{E}_P[\varphi S]$. Whether φ is unique depends on how rich the collection of scores is — a point made precise through the tangent space.

The terminological distinctions introduced in Chapter 10 can now be made precise:

- An **estimating function** is any function $U(O; \theta)$ whose root defines an estimator.
- An **asymptotic influence function** of an estimator is the function φ in its asymptotic expansion.
- A **(pathwise) influence function** of a functional is a $\varphi \in L_2^0(P)$ satisfying Equation C.1.

For a regular asymptotically linear estimator of a pathwise-differentiable functional, the estimator's asymptotic influence function is also a pathwise influence function of the functional. The relationship to estimating functions is looser: for an M-estimator solving $n^{-1} \sum_i U(O_i; \theta) = 0$, the standard expansion gives $\varphi(O) = -A^{-1} U(O; \theta_0)$ with $A = \mathbb{E}_P[\partial U(O; \theta_0)/\partial \theta^\top]$, so a generic estimating function U equals the influence function φ only after this normalization. The first two notions can be defined without reference to the model $\mathcal{P}$, while the third depends crucially on $\mathcal{P}$ through the collection of admissible submodels.

C.5 The Tangent Space

Definition: Tangent Space

The **tangent space** of $\mathcal{P}$ at P, denoted $\mathcal{T}_P(\mathcal{P})$ or simply $\mathcal{T}$, is the closed linear span in $L_2^0(P)$ of all scores of regular parametric submodels of $\mathcal{P}$ passing through P.

Two extreme cases are illustrative. **Nonparametric model.** If $\mathcal{P}$ contains all distributions on $\mathcal{O}$ satisfying mild regularity, then for any bounded $h \in L_2^0(P)$ the submodel $p_\varepsilon(o) \propto (1 + \varepsilon h(o))p(o)$ is regular for $|\varepsilon| < 1/\|h\|_\infty$, with score $S = h$. Since bounded mean-zero functions are dense in $L_2^0(P)$, taking closure gives $\mathcal{T} = L_2^0(P)$. **Fully parametric model.** If $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ with score components $S_{\theta_0,1}, \dots, S_{\theta_0,k}$, then $\mathcal{T} = \text{span}\{S_{\theta_0,1}, \dots, S_{\theta_0,k}\}$ is k-dimensional.

Definition: Nuisance Tangent Space

The **nuisance tangent space** of Ψ at P, denoted $\mathcal{T}_\eta$, is the closed linear span in $L_2^0(P)$ of scores of regular submodels along which the pathwise derivative of Ψ vanishes:

$$\mathcal{T}_\eta = \overline{\text{span}}\left\{ S \in \mathcal{T} : \frac{\partial}{\partial \varepsilon}\Psi(P_\varepsilon)\Big|_{\varepsilon=0} = 0,\ S \text{ the score of a regular submodel } \{P_\varepsilon\} \right\}.$$

A submodel with this property is called a **nuisance submodel**.

The nuisance tangent space is a closed subspace of $\mathcal{T}$, and hence of $L_2^0(P)$. Its orthogonal complement is taken in $L_2^0(P)$:

$$\mathcal{T}_\eta^\perp := \{f \in L_2^0(P) : \mathbb{E}_P[fg] = 0 \text{ for all } g \in \mathcal{T}_\eta\}. \tag{C.2}$$

The Hilbert-space decomposition $L_2^0(P) = \mathcal{T}_\eta \oplus \mathcal{T}_\eta^\perp$ holds automatically by the projection theorem and provides the geometric structure underlying semiparametric efficiency.

Proposition: Every Influence Function Lies in $\mathcal{T}_\eta^\perp$

Let Ψ be pathwise differentiable at P. Then every influence function $\varphi \in \text{IF}(\Psi, P)$ satisfies $\varphi \in \mathcal{T}_\eta^\perp$.

Proof

For any nuisance submodel $\{P_\varepsilon\}$ with score S_η, $\partial_\varepsilon \Psi(P_\varepsilon)|_0 = 0$ by definition of $\mathcal{T}_\eta$. Combined with Equation C.1: $0 = \mathbb{E}_P[\varphi S_\eta]$. This holds for every score S_η of a nuisance submodel; by linearity and L_2-continuity it extends to every element of the closed linear span $\mathcal{T}_\eta$. Hence $\varphi \perp \mathcal{T}_\eta$, i.e., $\varphi \in \mathcal{T}_\eta^\perp$. □

The distinguishing property of the efficient influence function is not membership in $\mathcal{T}_\eta^\perp$ (which is automatic) but rather that it is the unique influence function lying in the full tangent space $\mathcal{T}$.

Remark: Causal Example

For the ATE under consistency, conditional exchangeability, and positivity, the causal estimand is identified with the observed-data functional $\tau(P) = \mathbb{E}_P[\mu_1(X) - \mu_0(X)]$, $\mu_t(X) = \mathbb{E}_P[Y \mid T{=}t, X]$. In the nonparametric observed-data model for $O = (X, T, Y)$, $\mathcal{T} = L_2^0(P)$, and $\mathcal{T}_\eta$ is the closed linear span of scores of submodels along which the pathwise derivative of τ vanishes. Section Section C.7 carries out the construction in detail and shows that $\mathcal{T}_\eta^\perp$ is one-dimensional, spanned by the AIPW influence function; this is the formal content of the claim in Chapter 10 that φ^* is uniquely determined.

C.6 The Canonical Gradient and the Efficiency Bound

Every influence function lies in $\mathcal{T}_\eta^\perp$. What distinguishes a single influence function within this set? The answer is orthogonality to $\mathcal{T}^\perp$, equivalently membership in the full tangent space $\mathcal{T}$: among all influence functions, there is a unique one lying in $\mathcal{T}$, and it has the smallest variance.

Theorem: Canonical Gradient

Let Ψ be pathwise differentiable at P with non-empty influence function set $\mathrm{IF}(\Psi, P)$. Then:

(i) Any two influence functions differ by an element of $\mathcal{T}^\perp$: for $\varphi_1, \varphi_2 \in \mathrm{IF}(\Psi, P)$, $\varphi_1 - \varphi_2 \in \mathcal{T}^\perp$.
(ii) There exists a unique $\varphi^* \in \mathrm{IF}(\Psi, P)$ with $\varphi^* \in \mathcal{T}$, namely $\varphi^* = \Pi[\varphi \mid \mathcal{T}]$ for any $\varphi \in \mathrm{IF}(\Psi, P)$.
(iii) For every $\varphi \in \mathrm{IF}(\Psi, P)$: $\mathbb{E}_P[\varphi(O)^2] \geq \mathbb{E}_P[\varphi^*(O)^2]$, with equality iff $\varphi = \varphi^*$ in $L_2(P)$.

Proof

(i) For $\varphi_1, \varphi_2 \in \mathrm{IF}(\Psi, P)$ and every score S of a regular submodel, $\mathbb{E}_P[(\varphi_1 - \varphi_2)S] = \partial_\varepsilon \Psi|_0 - \partial_\varepsilon \Psi|_0 = 0$. By linearity and L_2-continuity, $\mathbb{E}_P[(\varphi_1 - \varphi_2)g] = 0$ for every $g \in \mathcal{T}$, i.e., $\varphi_1 - \varphi_2 \in \mathcal{T}^\perp$.
(ii) Fix any $\varphi \in \mathrm{IF}(\Psi, P)$ and set $\varphi^* := \Pi[\varphi \mid \mathcal{T}]$, so $\varphi - \varphi^* \in \mathcal{T}^\perp$. For any score $S \in \mathcal{T}$: $\mathbb{E}_P[\varphi^* S] = \mathbb{E}_P[\varphi S] - \mathbb{E}_P[(\varphi - \varphi^*)S] = \mathbb{E}_P[\varphi S] = \partial_\varepsilon \Psi(P_\varepsilon)|_0$, so $\varphi^* \in \mathrm{IF}(\Psi, P)$. Uniqueness: if $\tilde{\varphi}^* \in \mathrm{IF}(\Psi, P) \cap \mathcal{T}$, then by (i), $\varphi^* - \tilde{\varphi}^* \in \mathcal{T}^\perp \cap \mathcal{T} = \{0\}$.
(iii) Write $\varphi = \varphi^* + (\varphi - \varphi^*)$ with the two summands orthogonal in $L_2(P)$ (since $\varphi^* \in \mathcal{T}$ and $\varphi - \varphi^* \in \mathcal{T}^\perp$). The Pythagorean identity gives $\mathbb{E}_P[\varphi^2] = \mathbb{E}_P[(\varphi^*)^2] + \mathbb{E}_P[(\varphi - \varphi^*)^2] \geq \mathbb{E}_P[(\varphi^*)^2]$, with equality iff $\varphi = \varphi^*$. □

Definition: Efficient Influence Function and Efficiency Bound

The unique element $\varphi^* \in \mathrm{IF}(\Psi, P) \cap \mathcal{T}$ is called the **efficient influence function (EIF)**, or **canonical gradient**, of Ψ at P. Its variance $V^*(\Psi, P) = \mathbb{E}_P[\varphi^*(O)^2]$ is the **semiparametric efficiency bound** for estimating $\Psi(P)$ in the model $\mathcal{P}$.

Remark: Role of the Nuisance Tangent Space

Since every influence function already lies in $\mathcal{T}_\eta^\perp$, the EIF can equivalently be described as the unique influence function in $\mathcal{T} \cap \mathcal{T}_\eta^\perp$. This intersection contains exactly the directions in $\mathcal{T}$ that are orthogonal to nuisance perturbations — directions along which ψ genuinely moves. In practice, the EIF is often constructed by taking a candidate function in $\mathcal{T}$ (such as an unbiased estimating function for ψ) and subtracting its projection onto $\mathcal{T}_\eta$ to remove the nuisance components; this is the projection view exploited in Chapter 11.

Theorem: Asymptotic Efficiency Bound

Let $\hat{\psi}_n$ be a regular asymptotically linear estimator of $\Psi(P)$ in the model $\mathcal{P}$. Then $\mathrm{AVar}(\sqrt{n}(\hat{\psi}_n - \psi)) \geq V^*(\Psi, P)$, and the bound is achieved if and only if the influence function of $\hat{\psi}_n$ is φ^* in $L_2(P)$.

For regular asymptotically linear estimators, the variance bound follows directly from the Canonical Gradient Theorem (iii) applied to the estimator's influence function. The Hájek–Le Cam convolution theorem extends the lower-bound interpretation beyond the asymptotically linear class: for any regular estimator $\hat{\psi}_n$, $\sqrt{n}(\hat{\psi}_n - \psi) \rightsquigarrow Z + W$ where $Z \sim N(0, V^*)$ and W is independent of Z, so the variance bound persists by taking variances. A precise statement requires the local asymptotic normality framework of Vaart (1998, secs. 25.3–25.6).

Remark: Why the EIF Appears in Concrete Formulas

The Canonical Gradient Theorem explains why specific objects such as the AIPW influence function look "constructed" rather than guessed. In concrete causal examples one may start from an unbiased

estimating function such as the IPW score $U_{\text{IPW}}(O) = TY/\pi(X) - (1-T)Y/(1-\pi(X)) - \tau$. For the nonparametric ATE model, U_{IPW} already lies in $\mathcal{T} = L_2^0(P)$ and is unbiased when π is known, yet it fails to satisfy Equation C.1 for $\tau(P)$ when π is unknown — because perturbations of π contribute first-order terms to $\partial_\varepsilon \mathbb{E}_{P_\varepsilon}[U_{\text{IPW}}(O)]\big|_0$ that the EIF must absorb. The EIF is obtained by solving the pathwise-derivative equation directly, or equivalently by constructing the canonical gradient within $\mathcal{T} \cap \mathcal{T}_\eta^\perp$; for the nonparametric ATE this calculation produces the AIPW influence function.

C.7 Worked Example: The ATE Functional

This section makes the geometric machinery concrete by carrying out the EIF derivation for the ATE under the nonparametric observed-data model. The end product is the AIPW influence function. The value of the derivation lies in showing how it arises directly from the Canonical Gradient Theorem as the Riesz representer in $L_2^0(P)$ of the pathwise derivative of τ — recovering the AIPW formula from first principles.

Setup. The observed data are $O = (X, T, Y)$. Under consistency, conditional exchangeability, and positivity (Chapter 3), the ATE is identified with $\tau(P) = \mathbb{E}_P[\mu_1(X) - \mu_0(X)]$, $\mu_t(X) = \mathbb{E}_P[Y \mid T{=}t, X]$. Write $\pi(X) = P(T{=}1 \mid X)$. The model $\mathcal{P}$ is nonparametric: no restrictions are placed on the joint law of (X, T, Y) beyond positivity $0 < \pi(X) < 1$ a.s. Hence $\mathcal{T} = L_2^0(P)$.

Score factorization. The joint density factors as $p(x, t, y) = p_X(x) \cdot p_{T|X}(t \mid x) \cdot p_{Y|T,X}(y \mid t, x)$. Along any regular submodel the score decomposes additively:

$$S(O) = S_X(X) + S_T(T \mid X) + S_Y(Y \mid T, X),$$

with $\mathbb{E}_P[S_X(X)] = 0$, $\mathbb{E}_P[S_T(T \mid X) \mid X] = 0$, $\mathbb{E}_P[S_Y(Y \mid T, X) \mid T, X] = 0$. Define the closed subspaces of $L_2^0(P)$:

$$\mathcal{H}_X = \{a(X) : \mathbb{E}_P[a(X)] = 0\}, \quad \mathcal{H}_T = \{b(X)(T{-}\pi(X)) : b \in L_2(P_X)\}, \quad \mathcal{H}_Y = \{c(O) : \mathbb{E}_P[c(O) \mid T, X] = 0\}.$$

A short conditioning calculation shows these three subspaces are pairwise orthogonal in $L_2^0(P)$. For instance, for $a(X) \in \mathcal{H}_X$ and $b(X)(T - \pi(X)) \in \mathcal{H}_T$: $\mathbb{E}_P[a(X) \cdot b(X)(T - \pi(X))] = \mathbb{E}_P[a(X)b(X) \cdot \mathbb{E}_P\{T - \pi(X) \mid X\}] = 0$, since $\mathbb{E}_P[T \mid X] = \pi(X)$. The other two pairs follow analogously by conditioning on X and on (T, X) respectively. Joint spanning follows by writing any $g \in L_2^0(P)$ as a telescoping sum of conditional expectations and centering each piece. Hence:

$$L_2^0(P) = \mathcal{H}_X \oplus \mathcal{H}_T \oplus \mathcal{H}_Y. \tag{C.3}$$

The pathwise derivative. Differentiating $\tau(P_\varepsilon) = \int (\mu_{1,\varepsilon}(x) - \mu_{0,\varepsilon}(x))\, p_{X,\varepsilon}(x)\, d\mu(x)$ at $\varepsilon = 0$ and applying the product rule gives:

$$\frac{\partial}{\partial \varepsilon}\tau(P_\varepsilon)\Big|_0 = \underbrace{\int \{\mu_1(x) - \mu_0(x)\}\, S_X(x)\, p(x)\, d\mu(x)}_{(I)} + \underbrace{\int (\partial_\varepsilon \mu_{1,\varepsilon}(x) - \partial_\varepsilon \mu_{0,\varepsilon}(x))\big|_0 p(x)\, d\mu(x)}_{(II)}.$$

No S_T term appears: $\tau(P)$ is a functional of p_X and $p_{Y|T,X}$ only, so perturbations of the treatment law do not affect τ to first order. This already shows that $\mathcal{H}_T \subset \mathcal{T}_\eta$.

Term (I). Since $\mathbb{E}_P[S_X] = 0$, we may subtract any constant from $\mu_1(X) - \mu_0(X)$; the choice $\tau = \mathbb{E}_P[\mu_1(X) - \mu_0(X)]$ places the result in $\mathcal{H}_X$:

$$(I) = \mathbb{E}_P[\{\mu_1(X) - \mu_0(X) - \tau\}\, S_X(X)] = \langle \varphi_X, S_X \rangle_P, \quad \varphi_X(X) := \mu_1(X) - \mu_0(X) - \tau \in \mathcal{H}_X.$$

Term (II). Fix $t \in \{0, 1\}$ and compute:

$$\partial_\varepsilon \mu_{t,\varepsilon}(x)\big|_0 = \mathbb{E}_P[(Y - \mu_t(x))\, S_Y(Y \mid t, X) \mid T{=}t,\, X{=}x],$$

using $\mathbb{E}_P[S_Y(Y \mid t,x) \mid T{=}t, X{=}x] = 0$ to subtract $\mu_t(x)$. Write $\pi_t(x) = P(T{=}t \mid X{=}x)$. Using the identity $p(x)\,p(y \mid t,x) = p(x, T{=}t, y)/\pi_t(x)$ and applying for $t = 1$ and $t = 0$:

$$(II) = \mathbb{E}_P\left[\left\{\frac{T(Y-\mu_1(X))}{\pi(X)} - \frac{(1-T)(Y-\mu_0(X))}{1-\pi(X)}\right\} S_Y(Y \mid T, X)\right] = \langle \varphi_Y,\, S_Y\rangle_P,$$

with $\varphi_Y(O) \coloneqq T(Y-\mu_1(X))/\pi(X) - (1-T)(Y-\mu_0(X))/(1-\pi(X)) \in \mathcal{H}_Y$.

The canonical gradient. Combining and using the orthogonality of Equation C.3:

$$\frac{\partial}{\partial \varepsilon}\tau(P_\varepsilon)\Big|_0 = \langle \varphi^*,\, S\rangle_P, \qquad \varphi^*(O) \coloneqq \varphi_X(X) + \varphi_Y(O), \tag{C.4}$$

for every score $S = S_X + S_T + S_Y \in \mathcal{T} = L_2^0(P)$. By Definition Equation C.1, φ^* is an influence function of τ. By construction $\varphi^* \in \mathcal{H}_X \oplus \mathcal{H}_Y \subset L_2^0(P) = \mathcal{T}$, so φ^* is the canonical gradient. Writing it out explicitly:

$$\varphi^*(O) = \mu_1(X) - \mu_0(X) - \tau + \frac{T(Y-\mu_1(X))}{\pi(X)} - \frac{(1-T)(Y-\mu_0(X))}{1-\pi(X)}, \tag{C.5}$$

which is the AIPW influence function derived in Chapter 10.

Dimension of $\mathcal{T}_\eta^\perp$. The score-to-derivative map $\Lambda : \mathcal{T} \to \mathbb{R}$, $\Lambda(S) = \partial_\varepsilon \tau(P_\varepsilon)|_0$, is by Equation C.4 a non-zero continuous linear functional on $L_2^0(P)$ with Riesz representer φ^*. Its kernel is exactly $\mathcal{T}_\eta$, which by the codimension-and-uniqueness fact of Section Section C.1 has codimension one. Hence $\mathcal{T}_\eta^\perp = \operatorname{span}\{\varphi^*\}$ is one-dimensional, as asserted in the Causal Example remark above.

Reading off the efficiency bound. By the definition of the EIF, the semiparametric efficiency bound for estimating $\tau(P)$ in the nonparametric model is $V^*(\tau, P) = \mathbb{E}_P[\varphi^*(O)^2]$. Standard manipulations (iterated expectations on each summand of φ^* using the orthogonality of $\mathcal{H}_X, \mathcal{H}_Y$) decompose this as:

$$V^*(\tau, P) = \mathbb{E}_P\left[\frac{\sigma_1^2(X)}{\pi(X)} + \frac{\sigma_0^2(X)}{1-\pi(X)} + \{\mu_1(X) - \mu_0(X) - \tau\}^2\right],$$

with $\sigma_t^2(X) = \operatorname{Var}_P(Y \mid T{=}t, X)$ — the classical semiparametric variance bound for the ATE (Robins et al. 1994). By the Asymptotic Efficiency Bound Theorem, any regular asymptotically linear estimator of τ achieves this bound exactly when its influence function equals φ^* in $L_2(P)$, the analytic statement underpinning the asymptotic optimality of the AIPW estimator established in Chapter 11.

Bibliographic Notes

The modern formulation of pathwise differentiability and tangent spaces is developed in Bickel et al. (1993) and Vaart (1998, chap. 25); the latter is the standard reference for the convolution theorem and local asymptotic normality. Tsiatis (2006) gives a treatment oriented specifically toward missing data and causal inference, and is a natural companion to the material developed here.

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74 (1): 235–67.

Abadie, Alberto, and Guido W. Imbens. 2016. "Matching on the Estimated Propensity Score." *Econometrica* 84 (2): 781–807. https://doi.org/10.3982/ECTA11293.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2012. "Who Benefits from KIPP?" *Journal of Policy Analysis and Management* 31 (4): 837–60.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106 (4): 979–1014.

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61 (4): 962–73.

Baron, Reuben M., and David A. Kenny. 1986. "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51 (6): 1173–82.

Bartik, Timothy J. 1991. *Who Benefits from State and Local Economic Development Policies?* W. E. Upjohn Institute for Employment Research.

Berkson, Joseph. 1946. "Limitations of the Application of Fourfold Table Analysis to Hospital Data." *Biometrics Bulletin* 2 (3): 47–53.

Bezuidenhout, Dana, Sarah Forthal, Kara Rudolph, and Matthew R. Lamb. 2025. "Single World Intervention Graphs (SWIGs): A Practical Guide." *American Journal of Epidemiology* 194: 2047–52. https://doi.org/10.1093/aje/kwae353.

Bickel, Peter J., Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. 1993. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90 (430): 443–50.

Card, David. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, edited by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky. University of Toronto Press.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, et al. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–68.

Cinelli, Carlos, and Chad Hazlett. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Journal of the Royal Statistical Society, Series B* 82 (1): 39–67. https://doi.org/10.1111/rssb.12348.

Cochran, William G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24 (2): 295–313. https://doi.org/10.2307/2528036.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96 (1): 187–99. https://doi.org/10.1093/biomet/asn055.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programmes." *Journal of the American Statistical Association* 94 (448): 1053–62.

Deville, Jean-Claude, and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87 (418): 376–82.

Ding, Peng. 2024. *A First Course in Causal Inference*. CRC Press.

Ding, Peng, Avi Feller, and Luke Miratrix. 2016. "Randomization Inference for Treatment Effect Variation." *Journal of the Royal Statistical Society: Series B* 78 (3): 655–71.

Ding, Peng, and Tyler J. VanderWeele. 2016. "Sensitivity Analysis Without Assumptions." *Epidemiology* 27 (3): 368–77. https://doi.org/10.1097/EDE.0000000000000457.

Firth, David, and Karen E. Bennett. 1998. "Robust Models in Probability Sampling." *Journal of the Royal Statistical Society: Series B* 60 (1): 3–21.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik Instruments: What, When, Why, and How." *American Economic Review* 110 (8): 2586–624.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 (4): 1029–54. https://doi.org/10.2307/1912775.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46 (6): 1251–71.

Hernán, Miguel A., and James M. Robins. 2006. "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology* 17 (4): 360–72.

Hernán, Miguel A., and James M. Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25 (1): 51–71.

Imbens, Guido W., and Whitney K. Newey. 2009. "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity." *Econometrica* 77 (5): 1481–512.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Isaki, Cary T., and Wayne A. Fuller. 1982. "Survey Design Under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77 (377): 89–96.

Khan, Shakeeb, and Elie Tamer. 2010. "Irregular Identification, Support Conditions, and Inverse Weight Estimation." *Econometrica* 78 (6): 2021–42.

Kitagawa, Toru. 2015. "A Test for Instrument Validity." *Econometrica* 83 (5): 2043–63. https://doi.org/10.3982/ECTA11974.

Laan, Mark J. van der, and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer.

Laan, Mark J. van der, and Daniel Rubin. 2006. "Targeted Maximum Likelihood Learning." *The International Journal of Biostatistics* 2 (1): 1–40.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programmes with Experimental Data." *American Economic Review* 76 (4): 604–20.

Lauritzen, Steffen L. 1996. *Graphical Models*. Oxford University Press.

Levis, Alexander W., Edward H. Kennedy, and Luke Keele. 2024. "Nonparametric Identification and Efficient Estimation of Causal Effects with Instrumental Variables." *arXiv Preprint arXiv:2402.09332*.

Li, Xinran, and Peng Ding. 2017. "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference." *Journal of the American Statistical Association* 112 (520): 1759–69.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7 (1): 295–318.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review, Papers and Proceedings* 80 (2): 319–23.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. Springer. https://doi.org/10.1007/b97478.

Newey, Whitney K., and Richard J. Smith. 2004. "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators." *Econometrica* 72 (1): 219–55.

Neyman, Jerzy Splawa. 1923. "On the Application of Probability Theory to Agricultural Experiments:

Essay on Principles, Section 9." *Statistical Science* 5 (4): 465–72.

Richardson, Thomas S., and James M. Robins. 2014. *ACE Bounds; Single World Intervention Graphs (SWIGs) and Identification of Causal Effects*. University of Washington.

Robins, James M. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7 (9–12): 1393–512.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.

Rosenbaum, Paul R. 2002. *Observational Studies*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4757-3692-2.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.

Sargan, John D. 1958. "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica* 26 (3): 393–415. https://doi.org/10.2307/1907619.

Shachter, Ross D. 1998. "Bayes-Ball: The Rational Pastime (for Determining Irrelevance and Requisite Information in Belief Networks and Influence Diagrams)." *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 480–87.

Shpitser, Ilya, and Judea Pearl. 2006. "Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models." *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI)* 21: 1219–26.

Sobel, Michael E. 1982. "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models." *Sociological Methodology* 13: 290–312.

Tan, Zhiqiang. 2006. "A Distributional Approach for Causal Inference Using Propensity Scores." *Journal of the American Statistical Association* 101 (476): 1619–37. https://doi.org/10.1198/016214506000000023.

Tsiatis, Anastasios A. 2006. *Semiparametric Theory and Missing Data*. Springer.

Vaart, Aad W. van der. 1998. *Asymptotic Statistics*. Cambridge University Press.

VanderWeele, Tyler J., and Peng Ding. 2017. "Sensitivity Analysis in Observational Research: Introducing the E-Value." *Annals of Internal Medicine* 167 (4): 268–74. https://doi.org/10.7326/M16-2607.

Yang, Shu, Guido W. Imbens, Zhanglin Cui, Douglas E. Faries, and Zbigniew Kadziola. 2016. "Propensity Score Matching and Stratification in Observational Studies with Multi-Level Treatments." *Biometrics* 72: 1055–65.

Yang, Shu, and Yunshu Zhang. 2023. "Multiply Robust Matching Estimators of Average and Quantile Treatment Effects." *Scandinavian Journal of Statistics* 50: 235–65.

Zhao, Qingyuan, Dylan S. Small, and Bhaswar B. Bhattacharya. 2019. "Sensitivity Analysis for Inverse Probability Weighting Estimators via the Percentile Bootstrap." *Journal of the Royal Statistical Society, Series B* 81 (4): 735–61. https://doi.org/10.1111/rssb.12327.