Frontier Topics in Empirical Economics

Lecture Notes for a Doctoral Course

Zibin Huang

College of Business
Shanghai University of Finance and Economics

First draft, August 2026

Frontier Topics in Empirical Economics

Zibin Huang, College of Business, Shanghai University of Finance and Economics.

First draft, August 2026. Prepared from the author's lecture slides for the doctoral course of the same name.

Companion code and data:
https://github.com/hzbbb/Notes-on-Applied-Econometrics.
Practice examples were verified with Stata 17; data based figures were generated with Julia 1.12.

To my parents Wenlin Huang and Xiaoping Hong.

Foreword

[To be written.]

Preface

This textbook grew out of my teaching slides and notes for the second-year Ph.D. course at Shanghai University of Finance and Economics. The course spans fourteen weeks, with three classes per week. This course and textbook aim to (1) introduce the basic design-based research methods in economics to students who want to become applied economists; (2) introduce new techniques and tools in applied econometrics developed during the last decade; and (3) provide a brief introduction to the structural approach, encouraging students to further explore model-based research frameworks in the future.

To read this book, students should have completed their first-year courses in microeconomics, macroeconomics, and econometrics, and should understand the basic ideas of causal inference. It would be even better if they have some experience with structural, model-based methods, although this is not required. The content of this course is methodologically intensive, but its main focus is not on mathematical details. Rather, the most important goal is to understand how to use these methods correctly in empirical research.

To comprehensively develop students' abilities in both theoretical and applied aspects of empirical research, I include an application reading, a set of homework problems, and a practice coding exercise at the end of each chapter. The application reading recommends academic papers that use the methods introduced in the chapter. The homework problems aim to reinforce students' theoretical understanding of the material. The practice coding exercises give students an opportunity to write code themselves and apply the methods to real-world questions.

I divide the book into four parts comprising fourteen chapters.

In Part I, the book introduces the foundations of causal inference and basic statistical

tools. Chapter 1 discusses the basic ideas of causal inference under the potential outcomes framework and the relationship between causal inference and regression. Chapter 2 introduces non-parametric and semi-parametric methods, which provide important foundations for causal inference but are sometimes overlooked in the daily practice of applied economists. Chapter 3 introduces machine learning and model selection, showing how to determine model complexity by trading off bias and variance. Chapter 4 introduces graphical models for causal inference as an alternative to the traditional potential outcomes framework commonly used by economists.

In Part II, the book introduces one of the most popular and important methods in applied economics: instrumental variables (IV). Chapter 5 discusses the econometric foundations of IV, including two-stage least squares (2SLS), the local average treatment effect (LATE), and the generalized method of moments (GMM) framework. Chapter 6 discusses IV beyond the traditional LATE interpretation and shows how behavioral and structural models can be combined with IV. Chapter 7 discusses a specific type of IV that has become increasingly popular in recent years: the Bartik IV.

In Part III, the book introduces two other popular tools: difference-in-differences (DID) and regression discontinuity (RD). Chapter 8 introduces the basic tools used in panel data analysis, including fixed effects, DID, and synthetic control methods. Chapter 9 discusses more complicated and advanced topics in panel data analysis, including staggered DID with heterogeneous treatment effects and continuous treatments. Chapter 10 introduces the regression discontinuity design and its extension, the regression kink design.

In Part IV, the book discusses several other important topics. Chapter 11 discusses why and how to cluster standard errors. Chapter 12 introduces the fundamentals of discrete choice models, including their structure, identification, and the Logit model. Chapter 13 further discusses estimation and endogeneity in discrete choice models. Chapter 14 concludes the book by analyzing peer effects and spillovers in applied economics.

For readers using this book as a course text, I highly recommend following the chapters in order. In particular, readers should begin with Part I before moving on to specific topics in the subsequent parts, because the statistical tools introduced in Part I are repeatedly used throughout the rest of the book. After completing Part I, however, readers can also freely choose among the chapters in Parts II, III, and IV according to their interests. Compared with classic textbooks such as Angrist and Pischke (2009),

this book expands its scope to cover new tools and techniques that have been developed over the past decade, including graphical models, machine learning, Bartik IV, and more advanced methods for panel data analysis.

During the writing of this book, I would like to express my special thanks to Junjian Yi of Peking University. The applied econometrics reading group he organized inspired the design of this book. I would also like to thank my Ph.D. advisors, Ronni Pavan and Lisa Kahn. They gave me so many valuable suggestions during my doctoral studies and will always serve as role models throughout my academic career. I am truly grateful to Nese Yildiz, John Singleton, Kegon Tan, Gregorio Caetano, and Carolina Caetano for all the empirical techniques I learned from them during my time at the University of Rochester. I thank Dengke Chen for carefully going through all the teaching slides and notes in the first draft and providing me with many helpful suggestions. I thank Junsen Zhang for guiding me to choose to become a labor economist at the begining of my doctoral study. I would also like to thank Josh Angrist and Jörn-Steffen Pischke, whose magnificent textbook has illuminated my path toward becoming a researcher.

Last but not least, I would like to thank Claude Opus and Fable LLMs. They helped me reorganize my teaching slides and notes into a textbook and design many of the homework and practice problems. Without them, I could not have completed this work so efficiently. In this new era of AI, some people worry that we may eventually be replaced, while others may lose their motivation to study and develop their research in economics. I believe, however, that AI can help us reach places that we could never reach without it. It is the human mind that decides where to go, while AI helps us get there faster. I also firmly believe that students still need to work hard to learn econometrics and its technical details, because this knowledge is a prerequisite for developing good research taste. In other words, we need to know where to go before AI can take us all the way there.

As I write the final part of this preface, I suddenly find myself thinking back to my first day in the United States thirteen years ago: a young Chinese boy landing at Nashville International Airport. It was a brand new country for him, with no friends and no relatives. But it was also a brand new beginning in his life, filled with curiosity and ambition to explore the world. I want to say to that boy: Keep your happiness, and you will get there.

Zibin Huang, Wenzhou, August 2026

Contents

List of Figures

List of Tables

How to Use This Book

This book grew out of a fourteen week doctoral course. Each chapter corresponds to one week and has the same shape: the method and the reasoning behind it, one application paper read in detail, a set of homework problems, and a practice example with simulated data and Stata code that the reader can run and modify. The chapters build on one another in the order given, and the four parts group them by theme: Part I lays the foundations (potential outcomes, non-parametric methods, machine learning and model selection, and causal graphs); Part II is devoted to instrumental variables; Part III to panel data, difference-in-differences, and regression discontinuity; and Part IV to other topics that a modern empirical economist needs: inference with clustered data, discrete choice models, and peer effects.

Course map. The table gives the week by week schedule and the practice example attached to each chapter. The data and code are described in the Replication Guide at the end of the book.

Week	Chapter	Title	Practice example
1	1	Outline of Causal Inference	Estimating the ATE in a Randomized Controlled Trial
2	2	Non-parametric Methods	Non-parametric Regression in Stata
3	3	Machine Learning and Model Selection	LASSO and Ridge Regression in Stata
4	4	Directed Acyclic Graphs	Backdoor Adjustment in a Simulated DAG
5	5	Introduction to Instrumental Variables and the Endogeneity Problem	Returns to Schooling
6	6	IV beyond LATE	Weighted LATE with Two Binary Instruments
7	7	Bartik Instruments	Bartik IV with Simulated Shift-Share Data
8	8	Causal Inference with Panel Data I	A DID Simulation with Event Study
9	9	Causal Inference with Panel Data II	Negative Weights and DID_M in a Staggered Simulation
10	10	Regression Discontinuity Design	Sharp and Fuzzy RD in a Simulation
11	11	Standard Error Issues	The Moulton Problem in a Simulated Class Experiment
12	12	Discrete Choice Model I	Logit, Marginal Effects, and the IIA Property
13	13	Discrete Choice Model II	Berry Inversion, Linear IV for Demand, and IV-Probit
14	14	Peer Effect and Spillover	Mechanical, Spurious, and Fake Peer Effects

Prerequisites. A first year graduate sequence in econometrics (linear regression, asymptotic theory, maximum likelihood) and working familiarity with Stata. The discrete choice chapters also use basic numerical optimization.

Conventions. Theorems, propositions, definitions, and assumptions are numbered within sections (Theorem 5.2.1 is the first theorem of Section 5.2); equations and figures are numbered within chapters. Boxed text highlights key results (blue), practical

guidance (green), and warnings (red). Notation is collected in the list that follows.

Notation and Abbreviations

The chapters keep the notation of their source literatures, so a few symbols carry different meanings in different chapters; those cases are stated where they arise. The following conventions are common to the book.

Y_i, D_i, X_i	Outcome, binary treatment, and covariates of unit i
$Y_i(1)$, $Y_i(0)$	Potential outcomes under treatment and under no treatment
$\delta_i = Y_i(1) - Y_i(0)$	Individual treatment effect
ATE, ATT, ATU	Average treatment effect, on the treated, on the untreated
LATE, MTE	Local average treatment effect; marginal treatment effect
CATE	Conditional average treatment effect
CIA	Conditional independence assumption
$E[\cdot]$, $\text{Var}(\cdot)$, $\text{Cov}(\cdot, \cdot)$	Expectation, variance, covariance
$E[Y \mid X]$	Conditional expectation function (CEF)
$\perp\!\!\!\perp$	Statistical independence
$\xrightarrow{p}$, $\xrightarrow{d}$	Convergence in probability and in distribution
$\mathbf{1}(\cdot)$	Indicator function
$\bar{x}_g$, $\bar{S}_{-i,j}$	Group mean; leave-one-out group mean
$\hat{\theta}$, $\hat{\beta}$	Estimators
OLS, 2SLS, GMM, MLE	Ordinary least squares, two stage least squares, generalized method of moments, maximum likelihood
IV, RCT, DAG	Instrumental variable, randomized controlled trial, directed acyclic graph
DID, TWFE, FE, FD	Difference-in-differences, two way fixed effects, fixed effects, first differences
RD, RDD, RKD	Regression discontinuity (design), regression kink design
DCM, IIA, GEV	Discrete choice model, independence of irrelevant alternatives, generalized extreme value
T1EV	Type one extreme value distribution
LASSO, CV	Least absolute shrinkage and selection operator; cross-validation
SE, CI	Standard error, confidence interval

Part I

Foundations

Chapter 1

Outline of Causal Inference

1.1 What Is This Course About?

Empirical economics is built on a simple ambition: to understand how the world works by using data. But data are silent about causation on their own. They record correlations (associations between variables), but the question of whether A causes B, and by how much, requires more than a regression coefficient. This course is about the gap between *statistical association* and *causal inference*, and the modern toolkit that allows researchers to bridge it responsibly.

A central theme is what we might call the "regression monkey" problem. A regression monkey runs regressions without understanding why. They use whatever variables are available, add controls when reviewers complain, and report results without any sense of what variation in the data is actually doing the identification work. They confuse the tools of statistics with economic reasoning. This is not an economist; it is a bad statistician borrowing economic vocabulary.

What separates serious empirical work from mechanical computation? Three things. First, a researcher must understand the *logic* behind every estimator they use: why it targets a causal parameter under some assumptions, and what those assumptions mean economically. Second, they must know the menu of tools available (instrumental variables, difference-in-differences, regression discontinuity, matching) well enough to match the method to the research question. Third, they must use economic theory as a guide: theory tells us which variables belong in a model, which direction the bias

should run, and whether a result makes sense.

This chapter lays the conceptual foundation for the course. We begin with a motivating example that shows how a simple regression model is connected to a deeper structural model. We then introduce the two broad traditions in empirical economics, the structural (model-based) approach and the reduced-form (design-based) approach, and compare their relative strengths. The bulk of the chapter develops the potential outcome framework, the conditional expectation function and its relationship to regression, the problem of bad controls, and the logic of matching estimators.

1.2 Motivating Example: Female Labor Participation[1]

1.2.1 Two Levels of Empirical Analysis

To fix ideas, consider a classic question in labor economics: how does the wage rate affect the decision of women to enter the labor force? A researcher facing this question might immediately reach for a regression:

$$y_i = \beta_0 + \beta_1 w_i + u_i, \tag{1.1}$$

where $y_i \in \{0, 1\}$ is a labor-participation indicator, w_i is the offered wage, and u_i is the unobservable. Estimating (1.1) by OLS gives what is called a *linear probability model* (LPM). This is an example of a *reduced-form* regression: it directly relates an observable outcome to an observable treatment variable, without modeling the mechanism connecting them.

But the name "reduced-form" already hints at something deeper. The regression is the visible tip of an iceberg. Underneath it lies a behavioral model of individual decision-making that the regression *reduces* to a single equation. Understanding this connection is important, because it clarifies what β_1 is estimating, when the regression is valid, and what it cannot tell us.

[1]The author thanks Professor Chao Fu (University of Wisconsin-Madison) for introducing this example.

1.2.2 The Behavioral Model

Following the model due to Professor Chao Fu, suppose that female i chooses labor supply $l_i \in \{0, 1\}$ to maximize utility. Her problem is:

$$\max_{l_i} \; U_i(c_i,\; 1 - l_i) + \epsilon_{il}, \qquad \text{subject to} \quad c_i = w_i l_i, \tag{1.2}$$

where c_i is consumption, $1 - l_i$ is leisure, w_i is the wage, and ϵ_{il} is a taste shock that is unobservable to the researcher (though observed by the individual). The budget constraint $c_i = w_i l_i$ says that income equals wages times labor supply.

Because l_i is binary, we compare utilities under the two options. Individual i chooses to work ($l_i = 1$) if and only if the utility from working plus her taste shock for working exceeds the utility from not working:

$$U_i(w_i, 0) + \epsilon_{i1} \;\geq\; U_i(0, 1) + \epsilon_{i0}. \tag{1.3}$$

Rearranging, individual i works if and only if the difference in taste shocks falls below a threshold:

$$\epsilon_{i0} - \epsilon_{i1} \;\leq\; \epsilon^*(w_i), \qquad \text{where} \quad \epsilon^*(w_i) \equiv U_i(w_i, 0) - U_i(0, 1). \tag{1.4}$$

The threshold ϵ^* is the net utility gain from working (evaluated at the given wage) and it is increasing in w_i: higher wages make work more attractive, raising the threshold below which someone chooses to work.

1.2.3 The Working Probability and Two Research Approaches

Let F_ϵ denote the CDF of the shock difference $\epsilon_{i0} - \epsilon_{i1}$, which we take to be independent of w_i. The probability that individual i participates in the labor force, given her wage, is:

$$G(w) \;=\; \Pr(l = 1 \mid w) \;=\; \int_{-\infty}^{\epsilon^*(w)} dF_\epsilon \;=\; F_\epsilon(\epsilon^*(w))\,. \tag{1.5}$$

The function $G(w)$ is the object of interest: it tells us how labor supply responds to wages at the population level. Two very different empirical strategies can be used to learn about it.

The Reduced-Form / Design-Based Approach. The simplest approach is to assume that G is well approximated by a linear function:

$$G(w) \approx \beta_0 + \beta_1 w. \tag{1.6}$$

This gives the Linear Probability Model in (1.1). Regression then estimates $\beta_1 \approx \partial G/\partial w$, the effect of a one-unit increase in wage on the probability of participation. The key word is "reduced": this equation is derived (*reduced*) from the underlying structural model by imposing a linearity assumption and ignoring the internal structure. The researcher uses a research "design" (some source of exogenous variation in wages, such as a minimum wage change or a policy experiment) to estimate this parameter credibly. The method is called "design-based" because the identification strategy is the center of the analysis.

The Structural / Model-Based Approach. An alternative is to directly estimate the utility function U and the shock distribution F from the data. If we parameterize these objects, say, with parameter vectors Θ^U for the utility function and Θ^F for the shock distribution, then each observation contributes to the likelihood:

$$L(\Theta^U, \Theta^F;\ \text{data}) \ = \ \prod_{i=1}^{N} F_\epsilon(\epsilon^*(w_i))^{l_i} \left[1 - F_\epsilon(\epsilon^*(w_i))\right]^{1-l_i}, \tag{1.7}$$

which is maximized by MLE to recover Θ^U and Θ^F simultaneously. This approach retrieves the primitive parameters directly.

To see a concrete example, suppose utility is linear in the relevant arguments: $U_i(w_i l_i, 1 - l_i) = \alpha w_i l_i + \phi(1 - l_i)$, and the taste shocks ϵ_{i0} and ϵ_{i1} are i.i.d. Type I Extreme Value, so that their difference is logistic. Under these assumptions, the probability of working becomes the familiar logistic function, and (1.7) reduces to the logit likelihood:

$$L(\Theta^U, \Theta^F;\ \text{data}) = \prod_{i=1}^{N} \left(\frac{e^{\alpha w_i}}{e^{\alpha w_i} + e^{\phi}} \right)^{l_i} \times \left(\frac{e^{\phi}}{e^{\alpha w_i} + e^{\phi}} \right)^{1-l_i}. \tag{1.8}$$

One important issue in discrete choice models is that only differences in utility matter. That is, we can identify only the parameters that remain after taking differences in utility across alternative choices. This is because utility is intrinsically ordinal: observed choices are determined by the relative values of different alternatives rather than their

absolute utility levels. We will show more details on this discrete choice model/Logit model in Chapter 12. The structural approach thus *recovers the choice structure directly*, giving us estimates of the utility parameters α and ϕ rather than a reduced-form slope.

1.3 Two Approaches: Structural vs. Reduced-Form

1.3.1 Internal and External Validity

Before comparing the two approaches in detail, we need two foundational concepts that recur throughout applied work.

Definition 1.3.1 (Internal Validity)**.** A causal estimate is *internally valid* if it correctly identifies the causal effect within the specific population and setting under study. Internal validity is about getting the right answer for the particular context from which the data were drawn.

Definition 1.3.2 (External Validity)**.** A causal estimate is *externally valid* if the finding can be credibly extrapolated to populations, settings, or time periods other than those from which the data were drawn. External validity is about whether the result travels.

These two concepts are often in tension. An estimate that is internally valid for a specific experimental population may say little about a broader population. Conversely, a structural model that makes strong assumptions to achieve external generalizability may have low internal validity because those assumptions are unverifiable.

Heckman and Vytlacil (2007) usefully distinguish three layers of policy evaluation, which map onto this taxonomy. Consider China's One Child Policy (OCP) as an example:

1. **Evaluation of a historical intervention.** What was the causal impact of the OCP on China's fertility rate during the period it was in force? This is an internal question: we want a credible estimate for the specific historical context.
2. **Forecasting the effect of a similar intervention in a new environment.** What would the demographic impact be if China were to reinstate the one-child policy in 2023? This is an external question: we are extrapolating a previous policy from past conditions to a new time period and possibly a different economic

environment.

3. **Evaluation of a novel counterfactual never observed in history.** What would happen if the government mandated that all women give birth to at least one child? This is a further external question: we are considering a counterfactual that has never been observed anywhere. Thus, answering it requires a model of behavior; simply extrapolating from observed data is insufficient.

The first question can, in principle, be answered by a good design-based study. The second and third require some structural content, because we must extrapolate to situations outside the historical record. It is fairly simple to see that the difficulty of answering these three questions is increasing by order.

1.3.2 The Structural Approach: Advantages and Limitations

The structural approach aims to recover the primitive parameters of the economy (utility functions, production functions, market structures) and use them to simulate counterfactuals. Its primary advantages are:

Deeper economic content. Structural estimates recover the decision-making process itself. We learn not just *what* happens when a policy changes, but *why*: how agents respond given their preferences, constraints, and information. This economic depth makes results easier to interpret and harder to misuse.

External validity and resistance to the Lucas critique. The Lucas critique (1976) warns that reduced-form behavioral equations estimated under one policy regime will shift when the policy changes, because agents adjust their expectations and behavior. Structural models, by targeting deep parameters (preferences, technology) that are invariant to policy, are in principle immune to this critique. They can therefore support credible counterfactual analysis far from the observed data.

Limitations. The main cost is the need for strong, often untestable assumptions. The researcher must specify the functional form of utility, the distribution of unobserved shocks, the market structure, and so on. These choices are not determined by the data; they reflect modeling decisions. If the model is misspecified, the resulting estimates inherit that misspecification. Internal validity can therefore be low, in the sense that the structural estimates may not be robust to changes in modeling assumptions.

1.3.3 The Reduced-Form / Design-Based Approach: Advantages and Limitations

The design-based approach targets a specific marginal effect: what is the causal impact of treatment variable A on outcome B? It does not model the mechanism connecting them; the causal effect is, in a sense, a "black box." The approach builds its credibility from a research design, a source of exogenous variation in A that allows identification of its causal effect.

High internal validity. Because the approach relies on clean exogenous variation rather than modeling assumptions, its estimates are credible for the population from which the data are drawn, provided the identifying variation is genuinely exogenous.

Limitations. The effects identified are typically *local*: they apply to the individuals whose treatment status was affected by the source of variation used for identification. External validity requires additional assumptions. Moreover, because the mechanism is not recovered, the results are hard to use for out-of-sample counterfactuals. And the approach remains vulnerable to the Lucas critique when the researcher tries to use a local reduced-form estimate to predict the effect of a large-scale policy change.

1.3.4 The Target of This Book

The two approaches introduced above reflect two fundamental modes of scientific inquiry: deduction and induction. These represent two complementary ways in which humans seek to understand the world. The structural approach is grounded in deduction: it incorporates economic theory and subjective prior knowledge into a model and derives implications from general principles to specific cases. In contrast, the reduced-form approach is based on induction: it relies primarily on observed data and aims to uncover empirical relationships with minimal reliance on theoretical structure. This philosophy is often summarized by the phrase "letting the data speak" in statistics.

The debate between these two approaches is, at its core, a debate between two fundamental ways in which human beings seek to understand the world. This philosophical conflict dates back to thinkers such as Leibniz, Newton, Hume, and Kant, and has long been one of the central questions in epistemology. Neither approach is inherently superior; each has its own strengths and limitations. In practice, they often complement one another. A well-trained economist should be able to work with both approaches

and understand which framework is more appropriate for a given research question. A recent trend in empirical research is to combine these two approaches in order to leverage the strengths of both. In particular, researchers may first construct a structural model and then estimate its key parameters using design-based methods such as randomized controlled trials (RCTs) or difference-in-differences (DID). This strategy allows the model to be grounded in economic theory while ensuring that parameter estimates are credibly identified from the data.

This course focuses primarily on the reduced-form, or design-based, approach. Our goal is to provide students with a starting point for engaging in empirical economic research. In the later weeks of the course, we will also give a very brief introduction of the structural approach, with a particular focus on discrete choice models.

1.4 Potential Outcomes and the Randomized Controlled Trial

1.4.1 Correlation Is Not Causality

Consider the following data on health status and hospitalization drawn from a hypothetical survey: The data show that people who went to hospital have worse average

Group	Sample Size	Mean Health Status
Hospital	7,774	3.21
No hospital	90,049	3.93

Table 1.1: Health status by hospitalization status.

health status than those who did not. Should we conclude that going to hospital makes you sicker? Of course not. People go to hospital *because* they are already sick. The comparison is contaminated by the fact that hospitalization and health status are both driven by an underlying illness; the correlation between hospitalization and health status reflects selection into the hospital, not the causal effect of the hospital on health. This is a textbook example of *selection bias*: the treated and untreated groups are systematically different even before treatment.

This example crystallizes a fundamental challenge of empirical work. Observational

data record outcomes that are realized under actual treatment decisions. But to identify a causal effect, we would need to observe the same individual both treated and untreated, a logical impossibility. The potential outcome framework, which we now develop, provides a formal language for thinking about this problem.

1.4.2 The Potential Outcome Framework

The potential outcome framework, also known as the Rubin Causal Model, is the dominant formalism for causal inference in economics (Rubin, 1974). Its central objects are the *potential outcomes* associated with each possible treatment status.

Definition 1.4.1 (Potential Outcomes)**.** Let $D_i \in \{0, 1\}$ be a binary treatment indicator for individual i, and let Y_i be an observable outcome. Define:

- Y_{0i}: the **potential outcome** that individual i would experience if *not* treated ($D_i = 0$), regardless of their actual treatment status.
- Y_{1i}: the **potential outcome** that individual i would experience if *treated* ($D_i = 1$), regardless of their actual treatment status.

The key feature of these objects is that they are defined independently of the actual treatment received. Whether we observe Y_{0i} or Y_{1i} depends on D_i, but both potential outcomes are, in principle, well-defined properties of individual i. The connection between potential outcomes and the observed outcome is given by the *switching equation*:

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} = Y_{0i} + (Y_{1i} - Y_{0i})\, D_i. \tag{1.9}$$

The individual treatment effect for person i is the difference $Y_{1i} - Y_{0i}$: how much better or worse off would i be if treated versus untreated? This quantity is **fundamentally unobservable**. For any given individual, we observe the world either under treatment or under no treatment, never both simultaneously. This is the *fundamental problem of causal inference*: individual-level causal effects cannot be identified from a single observation. We can only work with averages.

1.4.3 Average Treatment Effects

Two population averages of the individual treatment effect are of primary interest.

Definition 1.4.2 (Average Treatment Effect and ATT)**.** The *Average Treatment Effect* (ATE) is:

$$\text{ATE} = E[Y_{1i} - Y_{0i}],$$

the expected treatment effect over a randomly drawn individual from the population. The *Average Treatment Effect on the Treated* (ATT) is:

$$\text{ATT} = E[Y_{1i} - Y_{0i} \mid D_i = 1],$$

the expected treatment effect conditional on actually receiving treatment.

The ATE and ATT differ when treated and untreated individuals have different potential outcomes on average, which is typical in observational settings where treatment is not random.

1.4.4 The Selection Bias Decomposition

In an observational study, the researcher typically observes the difference in mean outcomes between treated and untreated groups:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0].$$

Using the switching equation (1.9), this naive comparison can be decomposed as follows. First, expand using the definitions:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0].$$

Add and subtract $E[Y_{0i} \mid D_i = 1]$:

$$= \underbrace{E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1]}_{\text{ATT}} + \underbrace{E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]}_{\text{Selection bias}}. \tag{1.10}$$

This decomposition is fundamental. The first term is the ATT, the causal effect of treatment on the treated individuals. The second term is *selection bias*: the difference in untreated potential outcomes between the group that chose to be treated and the group that did not. In the hospitalization example, $E[Y_{0i} \mid D_i = 1]$ is the health status the hospitalized individuals would have had if they had not gone to hospital. This counterfactual health status is lower than the health status of those who actually chose

not to go to hospital, because people self-select into hospitalization based on being sick. Selection bias is therefore negative in that example: it causes the naive comparison to understate any beneficial effect of the hospital, and indeed to flip its sign entirely.

1.4.5 Randomization Eliminates Selection Bias

The elegance of a Randomized Controlled Trial (RCT) is that it eliminates selection bias by design. When treatment assignment is random (determined by a coin flip or lottery that is independent of any individual characteristic), the treatment indicator D_i is statistically independent of all individual attributes, including the potential outcomes:

$$D_i \perp\!\!\!\perp Y_{0i},\ Y_{1i}. \tag{1.11}$$

This is the *independence condition* for an ideal RCT. It implies that the two groups, treated and untreated, are on average identical in all respects, including their potential outcomes:

$$E[Y_{0i} \mid D_i = 1] = E[Y_{0i} \mid D_i = 0] = E[Y_{0i}].$$

The selection bias term in (1.10) therefore vanishes, and the simple difference-in-means identifies the average treatment effect:

$$\begin{aligned} E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] &= E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1] \\ &= E[Y_{1i} - Y_{0i} \mid D_i = 1] \\ &= E[Y_{1i} - Y_{0i}] = \text{ATE}. \end{aligned} \tag{1.12}$$

The last equality holds because, under randomization, D_i is independent of (Y_{0i}, Y_{1i}), so the distribution of potential outcomes is the same in the treated and untreated groups. ATT equals ATE under randomization.

This is why the RCT is the "gold standard" of causal inference. It does not require any assumptions about how treatment affects outcomes; it does not require any model of individual behavior. The only requirement is that treatment is genuinely randomly assigned.

1.5 Regression, the CEF, and Causal Inference

The potential outcome framework clarifies the *target* of causal inference. But in practice, most empirical work uses regression rather than a simple difference-in-means, because we rarely have clean randomization and we need to control for confounding variables. This section develops the relationship between regression, the Conditional Expectation Function (CEF), and causal inference, and asks: when does a regression coefficient carry a causal interpretation?

1.5.1 The Conditional Expectation Function

Definition 1.5.1 (Conditional Expectation Function)**.** Given a random outcome Y_i and a vector of conditioning variables X_i, the Conditional Expectation Function (CEF) is:

$$E[Y_i \mid X_i = x] = \int t \; f_Y(t \mid X_i = x) \; dt, \tag{1.13}$$

where $f_Y(\cdot \mid X_i = x)$ is the conditional density of Y_i given $X_i = x$.

The CEF is a population-level object. It answers the question: "in the population, what is the average value of Y among all individuals who share the covariate value $X_i = x$?" It is a description of the data-generating process, not an estimator.

Two important properties deserve emphasis.

Property 1: The CEF decomposition. We can always write:

$$Y_i = E[Y_i \mid X_i] + \varepsilon_i, \tag{1.14}$$

where the residual ε_i satisfies $E[\varepsilon_i \mid X_i] = 0$ by construction. This is not an assumption; it is an algebraic identity. It holds for any joint distribution of (Y_i, X_i) (as long as the CEF exists).

Property 2: The CEF is not necessarily causal. The CEF is the best statistical predictor of Y_i given X_i (in terms of minimizing MSE), but prediction is not causation. $E[Y_i \mid X_i]$ tells us what value of Y to expect for a person with covariates X_i; it does not tell us what would happen to that person's Y if we changed their X_i.

Theorem 1.5.1 (CEF Minimizes MSE)**.** *Let $m(X_i)$ be any measurable function of X_i.*

Then:

$$E[Y_i \mid X_i] = \underset{m(X_i)}{\arg\min}\ E\left[(Y_i - m(X_i))^2\right]. \tag{1.15}$$

That is, the CEF is the Minimum Mean Squared Error (MMSE) predictor of Y_i given X_i, within the class of all measurable functions.

Proof. Write $Y_i = E[Y_i \mid X_i] + \varepsilon_i$. For any function m:

$$\begin{aligned} E[(Y_i - m(X_i))^2] &= E[(E[Y_i \mid X_i] - m(X_i) + \varepsilon_i)^2] \\ &= E[(E[Y_i \mid X_i] - m(X_i))^2] + 2\,E[(E[Y_i \mid X_i] - m(X_i))\varepsilon_i] + E[\varepsilon_i^2]. \end{aligned}$$

The cross-product term is zero because $E[\varepsilon_i \mid X_i] = 0$ implies $E[(E[Y_i|X_i] - m(X_i))\varepsilon_i] = 0$. Therefore:

$$E[(Y_i - m(X_i))^2] = \underbrace{E[(E[Y_i \mid X_i] - m(X_i))^2]}_{\geq 0} + E[\varepsilon_i^2].$$

This is minimized by setting $m(X_i) = E[Y_i \mid X_i]$, which makes the first term zero. □

1.5.2 Linear Regression as Best Linear Predictor

The CEF minimizes MSE over *all* functions of X_i. Linear regression restricts attention to *linear* functions and finds the best one. Formally, the population regression coefficient β is defined as:

$$\beta = \underset{b \in \mathbb{R}^k}{\arg\min}\ E\left[(Y_i - X_i'b)^2\right]. \tag{1.16}$$

The first-order condition (the *moment condition* of regression) is obtained by differentiating with respect to b and setting to zero:

$$E[X_i(Y_i - X_i'\beta)] = 0. \tag{1.17}$$

Let $\varepsilon_i = Y_i - X_i'\beta$. Then (1.17) says $E[X_i\varepsilon_i] = 0$: the residual is uncorrelated with every regressor. Solving directly:

$$\beta = E[X_iX_i']^{-1}E[X_iY_i], \tag{1.18}$$

provided $E[X_i X_i']$ is invertible (the rank condition).

An important notational distinction must be kept clear. The *population* coefficient β in (1.18) is an identification concept: it exists in the population and characterizes the best linear predictor. The *OLS estimator* $\hat{\beta}_{\text{OLS}}$ is an estimation concept:

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y, \tag{1.19}$$

where X is the $n \times k$ matrix of observed data. The vector X_i is a random variable; X is its matrix of realizations. As sample size grows, $\hat{\beta}_{\text{OLS}} \to \beta$ under standard regularity conditions.

We usually use the OLS estimator to estimate the regression coefficient. Though, in principle, there can be other estimators. The sample OLS formula $\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y$ is also called the *sample analogue* of the population formula $\beta = E[X_i X_i']^{-1}E[X_i Y_i]$: it replaces population expectations with their sample counterparts.

1.5.3 When Does a Regression Coefficient Identify a Causal Effect?

A researcher can always compute a regression coefficient, provided the data satisfy the rank condition. But the fundamental question is: what is β *identifying*? Is it a causal parameter? The answer depends on whether the assumptions of the relevant causal model are satisfied. We consider two important cases.

Case 1: Unconditional Randomization. Suppose we have data from an experiment in which treatment D_i was randomly assigned, so that $D_i \perp\!\!\!\perp Y_{0i}, Y_{1i}$. Consider the regression:

$$Y_i = \alpha + \rho D_i + \varepsilon_i. \tag{1.20}$$

Under randomization (and with D_i binary), the CEF is linear in D_i, so the regression coefficient ρ equals the CEF slope. Computing directly:

$$\begin{aligned}\rho &= E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \\ &= E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0] \\ &= E[Y_{1i}] - E[Y_{0i}] \quad \text{(by independence)} \\ &= E[Y_{1i} - Y_{0i}] = \text{ATE}. \end{aligned} \tag{1.21}$$

Thus, under unconditional randomization, the regression coefficient on treatment directly estimates the ATE. Note that we also have $\rho = \text{ATT}$ under randomization, since the distribution of (Y_{0i}, Y_{1i}) is independent of D_i.

Case 2: Conditional Randomization (CIA / Selection on Observables). In observational studies, treatment is rarely unconditionally random. The *Conditional Independence Assumption* (CIA), also known as the *selection-on-observables* assumption, is the key identifying assumption for regression-based causal inference:

$$D_i \;\perp\!\!\!\perp\; Y_{0i},\, Y_{1i} \;\Big|\; X_i. \tag{1.22}$$

This says: conditional on the observed covariates X_i, treatment assignment is as good as random. The potential outcomes are independent of treatment once we account for X_i. Equivalently, after controlling for X_i, there is no remaining selection bias.

Under the CIA, the appropriate regression model includes controls:

$$Y_i = \alpha + \rho_r D_i + X_i'\gamma + \nu_i. \tag{1.23}$$

Homogeneous treatment effects. If the treatment effect is the same for all individuals (constant $\rho_r = Y_{1i} - Y_{0i}$), then under the CIA:

$$\begin{aligned}\rho_r &= E[Y_i \mid X_i, D_i = 1] - E[Y_i \mid X_i, D_i = 0] \\ &= E[Y_{1i} \mid X_i, D_i = 1] - E[Y_{0i} \mid X_i, D_i = 0] \\ &= E[Y_{1i} \mid X_i] - E[Y_{0i} \mid X_i] \quad \text{(by CIA)} \\ &= E[Y_{1i} - Y_{0i} \mid X_i] = E[Y_{1i} - Y_{0i}] = \text{ATE}. \end{aligned} \tag{1.24}$$

The regression coefficient identifies the treatment effect.

Heterogeneous treatment effects. When the treatment effect $\delta_x \equiv E[Y_{1i} - Y_{0i} \mid X_i = x]$ varies across individuals with different covariates, the analysis is more subtle. Define the within-cell treatment effect as:

$$\delta_x = E[Y_i \mid X_i = x, D_i = 1] - E[Y_i \mid X_i = x, D_i = 0],$$

and define the within-cell treatment variance:

$$\sigma_D^2(X_i) \equiv E\Big[(D_i - E[D_i \mid X_i])^2 \;\Big|\; X_i\Big].$$

Then, as shown in Angrist and Pischke (2009, Chapter 3.3.1), the regression coefficient from (1.23) is:

$$\rho_r = \frac{E[\sigma_D^2(X_i)\,\delta_x]}{E[\sigma_D^2(X_i)]}. \tag{1.25}$$

This is a **treatment-variance weighted average** of the group-level ATEs δ_x. The weight assigned to each cell x is proportional to the within-cell variance of treatment. Cells where all units are treated (or all are untreated) contribute zero weight because they contain no variation in treatment to exploit. Cells with roughly equal numbers of treated and untreated units, where D_i has the highest variance, contribute the most.

To understand the intuition behind this result: OLS is projecting Y_i onto D_i after partialing out X_i. The variation in D_i that regression uses is the within-cell deviation $D_i - E[D_i \mid X_i]$. Cells with more of this variation naturally receive more weight. When treatment effects are homogeneous, this weighting is harmless. When they are heterogeneous, the OLS coefficient is a weighted average of local treatment effects, with weights determined by the distribution of treatment propensity, not by the distribution of the population.

1.5.4 The Hierarchy of Assumptions

The difference between regression, the CEF, and a causal model lies in the strength of the definition. Specifically, they are defined over different assumptions of the residual e in the model $Y = f(D) + e$:

- **Linear regression** by construction requires only that the residual is *uncorrelated* with the regressors: $E[De] = 0$. This is the weakest condition.

- **The CEF** by construction requires *mean independence*: $E[e \mid D] = 0$. This is stronger than uncorrelatedness. It says the expected residual is zero not just on average but for every value of D.
- **A causal model under the CIA** by construction requires *full independence*: $e \perp\!\!\!\perp D$, equivalently $D_i \perp\!\!\!\perp Y_{0i}, Y_{1i}$. This is the strongest condition.

Hierarchy of Assumptions

$$\underbrace{e \perp\!\!\!\perp D}_{\text{Causal model (CIA)}} \succ \underbrace{E[e \mid D] = 0}_{\text{CEF}} \succ \underbrace{E[De] = 0}_{\text{Linear regression}}$$

Each condition implies all weaker conditions below it, but not conversely.

What is the relationship between the CEF and linear regression? When the CEF is itself a linear function of X_i, then the two coincide exactly: linear regression recovers the CEF. When the CEF is non-linear, linear regression still has a useful interpretation: it is the best linear approximation to the CEF, in the sense that it minimizes the expected squared error $E[(E[Y_i \mid X_i] - X_i'\beta)^2]$ over all linear functions. This means regression is always meaningful as a prediction tool, even if the CEF is non-linear.

It is also worth noting a useful special case: when D is a binary variable (a dummy), linear regression with only D as a regressor is identical to the CEF. This is because the CEF of Y on a binary D must be a linear function of D: there are only two values of D, so any function of D is linear in D.

Summary: Regression and Causal Inference

- The CEF is the best predictor of Y given X over all functions.
- Linear regression is the best *linear* predictor; it is the best linear approximation to the CEF even if the CEF is non-linear.
- A regression coefficient is a causal effect only when the CIA holds.
- Under CIA with homogeneous treatment effects, the regression coefficient equals the ATE.
- Under CIA with heterogeneous treatment effects, the regression coefficient is a treatment-variance weighted average of group-level ATEs, a meaningful but not necessarily population-representative causal parameter.

1.6 Simpson's Paradox, Omitted Variable Bias, and Bad Controls

The CIA is the most important assumption we need in causal inference. A natural question then arises: is it always better to control for more variables? The answer is no. What variables should we control for, and what variables should we not? In this section, we briefly discuss this issue.

1.6.1 Simpson's Paradox

A recurring source of confusion in empirical work is the phenomenon known as Simpson's Paradox: an association between treatment and outcome that reverses direction when a third variable is controlled for. Consider the following data on two COVID-19 treatments:

Treatment	Mild Condition	Severe Condition	Overall
A	15% (210/1400)	30% (30/100)	16% (240/1500)
B	10% (5/50)	20% (100/500)	19% (105/550)

Table 1.2: Death rates by treatment and patient condition (COVID-19 example).

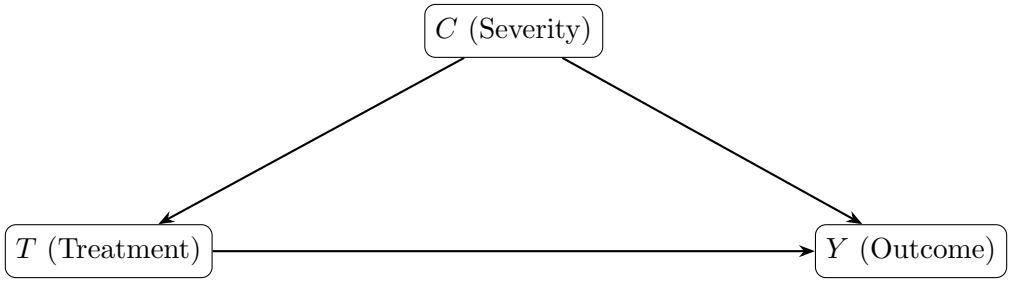
Looking at overall death rates, treatment A appears better: 16% versus 19%. Yet within each condition group, treatment A is worse: 15% vs. 10% for mild patients, and 30% vs. 20% for severe patients. How can A be simultaneously worse in every subgroup but better overall?

The resolution lies in the composition of the groups. Treatment A is more associated with mild patients (1400 out of 1500), who have low baseline death rates. Treatment B is more associated with severe patients (500 out of 550), who have high baseline death rates.

Which comparison is correct? This is a question equivalent to whether we should control for mild/severe condition when regressing death rate on treatment. The answer depends on the causal structure among these variables.

1.6.2 Confounders: When to Control

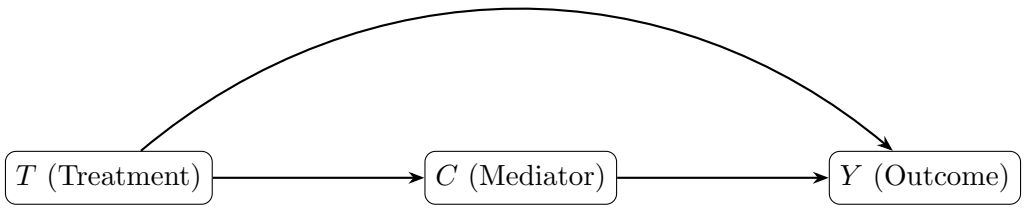
Case 1: The covariate C is a cause of treatment T. Suppose condition severity C affects both the treatment a patient receives T and the health outcome Y: the arrows run $C \to T$ and $C \to Y$, with $T \to Y$ as well. In this case, C is a *confounder*: it is a pre-existing characteristic that affects both treatment receipt and the outcome. If we compare treated and untreated patients without controlling for C, we are comparing groups that differed in C even before treatment was assigned. The difference in outcomes picks up not just the effect of T but also the effect of C; this is *Omitted Variable Bias*.



Case 1: C is a confounder. Must control for C.

The solution is to control for C: compare treated and untreated patients within each condition group. This is what the within-group comparisons in the table above show. Controlling for severity, treatment B is consistently better, and this comparison correctly isolates the causal effect of treatment.

Case 2: Treatment T is a cause of the covariate C. Now suppose instead that the treatment T itself causes the condition status C: for example, T may cause an immune response that leads to a "severe" classification on some post-treatment health index. The causal graph now has arrows $T \to C$ and $T \to Y$ (direct effect) and $C \to Y$ (mediated effect through C). In this case, C is a *mediator*: it lies on the causal pathway from T to Y.



*Case 2: C is a mediator. Do **not** control for C.*

If we control for C in this setting, we are holding constant something that treatment itself caused. We are effectively asking: "what is the effect of T on Y, holding C fixed?", but C cannot be held fixed independently of T, since T determines C. Controlling for C blocks the indirect causal channel from T to Y running through C, and the regression now captures only the direct effect of T on Y (not through C). If the goal is to estimate the total causal effect of T on Y, this is *wrong*. This is the **Bad Control Problem**.

Rule of Thumb: Pre-determined vs. Post-determined Variables

Pre-determined variables, those that are determined before the treatment is assigned and are not themselves caused by treatment, are generally safe to control for. Controlling for pre-determined confounders helps satisfy the CIA.

Post-determined variables, those that are caused by the treatment, or that happen after treatment assignment, should generally *not* be controlled for. Controlling for a mediator blocks the causal channel and introduces bias. Never control for a variable that lies on the causal pathway from treatment to outcome.

Let us apply this rule to three examples:

(a) Y = wage, D = education, X = natural ability. Natural ability is determined at birth, before any educational decision. It affects both education (more able individuals get more schooling) and wages (more able workers are more productive). X is a pre-determined confounder: we *should* control for it. Failing to do so overstates the return to education.

(b) Y = wage, D = education, X = labor participation decision. Labor market participation is a post-education choice: it happens *after* and is partly *caused by* D (education affects the decision to participate). X is post-determined: we should *not* control for it. Conditioning on labor participation selects a non-random subset of workers and introduces sample selection bias.

(c) $Y = \text{GDP}_{t+1}$, D = R&D expenditure at time t, X = trade volume at time t+1. Trade volume at $t+1$ is post-determined: it is realized after the R&D investment at t and may itself be a consequence of innovation. We should *not* control for it.

The DAG framework, introduced formally in Chapter 4, provides a systematic and rigorous tool for resolving such questions in general settings.

1.7 Matching

1.7.1 The Idea of Matching

Regression is only one way to use the CIA to estimate causal effects. A conceptually simpler and non-parametric alternative is *matching*. The basic idea is direct: find, for each treated unit, a control unit with *identical* observable characteristics X_i, and compare their outcomes. Because the two units have the same X_i, any difference in outcomes can be attributed to treatment alone (under the CIA), not to differences in the covariates.

Matching is attractive because it is non-parametric: it makes no functional form assumption about how X_i enters the outcome equation. It also has a simple and transparent interpretation: we are computing a within-cell average of the treatment effect, then averaging those within-cell effects across cells.

1.7.2 Formal Derivation of the Matching Estimator

Under the CIA ($Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i \mid X_i$), we can derive the matching estimators for ATT and ATE. For the ATT:

$$\begin{aligned}
\delta_{\text{ATT}} &= E[Y_{1i} - Y_{0i} \mid D_i = 1] \\
&= E\Big[E[Y_{1i} - Y_{0i} \mid X_i, D_i = 1] \,\Big|\, D_i = 1\Big] \quad \text{(law of iterated expectations)} \\
&= E\Big[E[Y_{1i} \mid X_i, D_i = 1] - E[Y_{0i} \mid X_i, D_i = 1] \,\Big|\, D_i = 1\Big] \\
&= E\Big[E[Y_i \mid X_i, D_i = 1] - E[Y_i \mid X_i, D_i = 0] \,\Big|\, D_i = 1\Big]. \qquad (1.26)
\end{aligned}$$

In the last step, we used the CIA: conditional on X_i, the treatment group and the control group have the same untreated potential outcome Y_{0i}, so $E[Y_{0i} \mid X_i, D_i = 1] = E[Y_{0i} \mid X_i, D_i = 0] = E[Y_i \mid X_i, D_i = 0]$. Defining the within-cell ATT as:

$$\delta_x = E[Y_i \mid X_i = x, D_i = 1] - E[Y_i \mid X_i = x, D_i = 0],$$

the ATT becomes $\delta_{\text{ATT}} = E[\delta_x \mid D_i = 1]$, estimated by the sample analog:

$$\hat{\delta}_{\text{ATT}} = \sum_x \hat{\delta}_x \, \hat{P}(X_i = x \mid D_i = 1). \qquad (1.27)$$

The ATT matching estimator weights each cell by the fraction of treated units that fall in that cell. Similarly, the ATE matching estimator weights by the unconditional share of each cell in the population:

$$\hat{\delta}_{\text{ATE}} = \sum_x \hat{\delta}_x \, \hat{P}(X_i = x). \tag{1.28}$$

1.7.3 Matching, Regression, and Weighting

A striking insight from Angrist and Pischke (2009) is that regression is itself a particular weighted matching estimator. All three estimators compute within- cell treatment effects $\hat{\delta}_x$ and then aggregate them, but they differ in the weights used:

Estimator	Formula	Weighting Principle
Matching (ATT)	$\sum_x \hat{\delta}_x \, \hat{P}(X_i = x \mid D_i = 1)$	Treated units' covariate distribution
Matching (ATE)	$\sum_x \hat{\delta}_x \, \hat{P}(X_i = x)$	Overall covariate distribution
Regression	$\frac{\sum_x \hat{\sigma}^2_D(x) \, \hat{\delta}_x \, \hat{P}(X_i = x)}{\sum_x \hat{\sigma}^2_D(x) \, \hat{P}(X_i = x)}$	Within-cell treatment variance

Table 1.3: The three estimators as weighted averages of cell-level treatment effects $\hat{\delta}_x$.

The ATT matching estimator asks: "on average, what is the treatment effect for a randomly drawn treated individual?", so it weights by the distribution of covariates among the treated. The ATE matching estimator asks: "on average, what is the treatment effect for a randomly drawn individual from the whole population?", so it weights by the population distribution. The regression estimator up-weights cells where treatment variance is high (where roughly half the units are treated) and down-weights cells where almost everyone is treated or untreated.

When treatment effects are homogeneous (constant $\delta_x = \delta$ for all x), all three estimators identify the same parameter δ. When treatment effects are heterogeneous, they identify different weighted averages, and the choice of estimator should reflect the policy question of interest.

1.8 Propensity Score Matching

1.8.1 The Curse of Dimensionality

Matching on covariates X_i is straightforward when X_i is a scalar or a small discrete vector: we simply find pairs of treated and untreated units with the same value of X_i and compare them. The method runs into serious difficulties, however, when X_i is high-dimensional.

Consider estimating the college wage premium while satisfying the CIA. The covariates needed to make D_i (college attendance) conditionally independent of potential wages might include gender, race, nationality, birth weight, IQ, parental income, parental education, and many more. With ten such covariates, each taking many values, the number of possible combinations of X_i is enormous. In a dataset of even 10,000 observations, most covariate cells will contain zero or one observation. We cannot compute $\hat{\delta}_x$ because we cannot find treated and untreated units with the same X_i. This is the *curse of dimensionality*: the data become too sparse to support matching as the dimension of X_i grows.

Regression side-steps this problem by imposing linearity: instead of matching within each cell, it extrapolates using the linear functional form. But matching is more flexible and transparent. Is there a way to do non-parametric matching even with high-dimensional covariates? The answer is yes, and it is provided by the Propensity Score Theorem.

1.8.2 The Propensity Score Theorem

Definition 1.8.1 (Propensity Score)**.** The *propensity score* is the conditional probability of receiving treatment given covariates:

$$P(X_i) = \Pr(D_i = 1 \mid X_i). \tag{1.29}$$

The propensity score is a single number that summarizes the relevant information in X_i for the purpose of matching. Its key property is given by the following theorem.

Theorem 1.8.1 (Propensity Score Theorem, Rosenbaum and Rubin 1983)**.** *Assume:*

(i) ***CIA****:* $Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i \mid X_i$.

(ii) ***Overlap****:* $0 < P(D_i = 1 \mid X_i) < 1$ *for all* X_i*.*

Then $Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i \mid P(X_i)$*.*

Proof sketch. It suffices to show $\Pr(D_i = 1 \mid Y_{1i}, Y_{0i}, P(X_i)) = P(X_i)$. By iterated expectations and CIA:

$$\begin{aligned} E[D_i \mid Y_{1i}, Y_{0i}, P(X_i)] &= E\Big[E[D_i \mid Y_{1i}, Y_{0i}, X_i] \,\Big|\, Y_{1i}, Y_{0i}, P(X_i)\Big] \\ &= E\Big[E[D_i \mid X_i] \,\Big|\, Y_{1i}, Y_{0i}, P(X_i)\Big] \quad \text{(by CIA)} \\ &= E\Big[P(X_i) \,\Big|\, Y_{1i}, Y_{0i}, P(X_i)\Big] \\ &= P(X_i). \end{aligned}$$

Since D_i is binary, $E[D_i \mid Y_{1i}, Y_{0i}, P(X_i)] = P(X_i)$ implies $D_i \perp\!\!\!\perp (Y_{1i}, Y_{0i}) \mid P(X_i)$. □

The theorem says: if we condition on the scalar propensity score $P(X_i)$ rather than the full vector X_i, the CIA is preserved. This is powerful: instead of matching on all the dimensions of X_i (infeasible when X_i is high-dimensional), it suffices to match on the single scalar $P(X_i)$. The curse of dimensionality is resolved by collapsing X_i into a one-dimensional sufficient statistic for selection into treatment.

1.8.3 Implementing PSM

In practice, the propensity score $P(X_i)$ is unknown and must be estimated. A typical PSM implementation proceeds in two steps:

1. **Estimate the propensity score.** Regress D_i on X_i using a binary model; logistic regression (logit) is the most common choice, though probit or even the linear probability model are also used. Obtain predicted probabilities $\hat{P}(X_i)$.
2. **Match and compare.** For each treated unit, find one or more control units with similar values of $\hat{P}(X_i)$, either via nearest-neighbor matching, kernel matching, or stratification on $\hat{P}(X_i)$. Average the within-match differences in Y_i to estimate ATT or ATE.

1.8.4 PSM Compared with Regression

Regression and PSM are both valid estimators under the CIA, but they have different properties:

- Regression does not suffer from the curse of dimensionality because the linearity assumption acts as a regularizer: it extrapolates smoothly across covariate cells using the fitted linear surface.
- PSM avoids functional form assumptions about the outcome equation, at the cost of requiring a model for the treatment probability. It also decreases estimation efficiency, which reflects a core concept in statistics: bias-variance tradeoff. We will discuss this in Chapter 3.
- A natural compromise is to run a regression of Y_i on D_i and $\hat{P}(X_i)$ (controlling for the propensity score rather than all covariates individually). This simple compromise, which controls for the estimated propensity score, provides some protection against misspecification.
- In practice, Angrist and Pischke (2009) argue that regression generally dominates PSM because the implementation of PSM involves discretionary choices (which estimator for $P(X_i)$? how many neighbors? what bandwidth?) that are not standardized and can affect results.

Critical Warning: PSM Cannot Solve Endogeneity Problem

Both regression and PSM assume CIA. They reduce the *dimensionality* of the matching problem, but they do not relax the identifying assumption. If there are unobserved variables that affect both treatment assignment and the outcome (that is, if selection is not fully explained by the observed X_i), then neither OLS nor PSM identifies a causal effect.

PSM CANNOT solve the endogeneity problem. It is a method for implementing the CIA efficiently, not a solution to violations of the CIA. If you cannot defend the CIA on substantive grounds, neither regression nor PSM will save you.

Homework Problems

1. **Weighted average under RCT.** Consider the expression for the regression coefficient under heterogeneous treatment effects:

$$\rho_r = \frac{E[\sigma_D^2(X_i)\,\delta_x]}{E[\sigma_D^2(X_i)]}.$$

   Show what this expression reduces to when unconditional independence holds, i.e., when $D_i \perp\!\!\!\perp Y_{1i}, Y_{0i}$ (as in an RCT without covariates). Interpret the result.

2. **Selection bias in observational data.** The data below show test scores for students who attended a tutoring program ($D_i = 1$) and those who did not ($D_i = 0$):

Group	n	Mean score
Tutored	200	75
Not tutored	800	70

   (a) What is the naive estimate of the effect of tutoring?
   (b) Students who sign up for tutoring are likely to be more motivated. Write down the selection bias formally using the potential outcome notation. What is the direction of the bias?
   (c) Describe an ideal RCT design that would identify the causal effect of tutoring.

Practice Example: Estimating the ATE in a Randomized Controlled Trial

Economic Setup

A government runs a job training program and randomly assigns 1,000 unemployed workers to either participate in the program ($D_i = 1$) or remain in the control group ($D_i = 0$). Monthly wages are measured six months after the program ends. Three pre-treatment characteristics are also recorded: age (years), education (years of schooling), and gender.

The goal is to estimate the *average treatment effect* (ATE) of the program on monthly

wages. This example illustrates the main message of Chapter 1: under random assignment, a simple OLS regression of the outcome on the treatment indicator identifies the ATE.

Data-Generating Process

The true model used to simulate the data is:

$$\begin{aligned} \text{wage}_i &= 1500 + 500 \cdot D_i + 80 \cdot \text{educ}_i + 30 \cdot (\text{age}_i - 35) - 200 \cdot \text{female}_i + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}(0,\ 300^2), \end{aligned} \tag{1.30}$$

with covariates drawn independently: age_i and educ_i are integer valued and approximately uniform on $[18, 60]$ and $[8, 20]$ respectively, and $\text{female}_i \sim \text{Bernoulli}(0.5)$.

The **true ATE is \$500 per month**. Treatment D_i is drawn independently of all covariates and the error term, so:

$$D_i \ \perp\!\!\!\perp \ (Y_{0i},\, Y_{1i},\, \text{age}_i,\, \text{educ}_i,\, \text{female}_i).$$

Regression Specifications

We estimate the ATE with two OLS specifications:

$$(1) \quad \text{wage}_i = \alpha + \rho\, D_i + e_i \qquad \text{(Simple OLS, no controls)} \tag{1.31}$$

$$(2) \quad \text{wage}_i = \alpha + \rho\, D_i + \beta_1\, \text{age}_i + \beta_2\, \text{educ}_i + \beta_3\, \text{female}_i + v_i \qquad \text{(OLS with controls)} \tag{1.32}$$

Under randomization, the coefficient ρ in specification (1) equals the ATE by the selection-bias decomposition: since $D_i \perp\!\!\!\perp Y_{0i}$, the selection bias term $E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0] = 0$, and the naive difference-in-means recovers the ATT, which equals the ATE.

Adding controls in specification (2) does not change the point estimate of ρ (because D_i is already independent of the controls), but it absorbs the variance in wages attributable to age, education, and gender, reducing the residual variance $\text{Var}(v_i)$ and therefore

shrinking the standard error of $\hat{\rho}$.

Expected Results

Table 1.4 presents the expected regression output (representative sample results; exact values will vary across simulation runs with different seeds).

	(1) Simple OLS	(2) With Controls
D_i (job training)	≈ 500	≈ 500
	(SE ≈ 34)	(SE ≈ 19)
age_i	—	≈ 30
educ_i	—	≈ 80
female_i	—	≈ -200
N	1,000	1,000
R^2	low	higher

Table 1.4: Expected regression results. True ATE = \$500/month. SE reported in parentheses. Coefficients on controls recover their true values from the DGP (1.30).

Key Learning Points

Running this example teaches three lessons that mirror the theoretical discussion in Sections 1.4 and 1.5.

1. **Identification under randomization.** The coefficient on D_i in specification (1) is approximately \$500, confirming that simple OLS identifies the ATE when treatment is randomly assigned. There is no need for controls to get an unbiased estimate.

2. **Invariance of the estimate to controls.** The coefficient on D_i does not change materially between specifications (1) and (2). This is the expected behavior under a genuine RCT: because D_i is independent of the covariates, adding them does not affect what variation in D_i is doing the work. A large change in the coefficient when controls are added would be a warning sign that the treatment was not truly random.

3. **Efficiency gain from pre-treatment controls.** The standard error of $\hat{\rho}$ falls

substantially in specification (2). Age, education, and gender together explain a large share of the variance in wages. Removing this variance from the residual tightens the confidence interval on the treatment effect without introducing any bias.

Balance Check

Before running regressions, it is good practice to verify that randomization produced balanced groups. A balance table should report means of pre-treatment covariates by treatment status, together with p-values from t-tests for equality of means. Under a genuine RCT with $N = 1{,}000$ and a 50/50 split, we expect all p-values to exceed 0.1, indicating no statistically significant differences between treated and control groups before treatment.

Files

- **Replication code**: `src/chapter 1/ch1_rct.do`
  Generates the dataset, runs the balance check, estimates both regressions, and prints a comparison table. Set the `global root` path at the top of the file before running.
- **Simulated data**: `data/chapter 1/ch1_rct.dta`
  Created automatically by running `ch1_rct.do` with `set seed 20250912`. Contains 1,000 observations and five variables: `wage`, `D`, `age`, `educ`, `female`.

Chapter 2

Non-parametric Methods

2.1 Introduction

In Chapter 1 we saw that the potential outcome framework and OLS regression can identify causal effects under certain assumptions. Throughout that chapter, the implicit assumption was that the conditional expectation function (CEF) is *linear* in the covariates. This chapter asks: what if linearity fails? Can we estimate the CEF directly from data, without imposing any functional form?

2.1.1 Why Go Beyond Parametric Models?

Consider the two most common parametric models in econometrics.

Linear regression assumes $y = X'\beta + e$, $e \sim N(0, \sigma^2)$. It is the workhorse of applied work: easy to estimate, easy to interpret, and often a reasonable approximation. But "often" is not "always."

Why do we assume the CEF is linear? Mostly for convenience. Why not $y = \beta x^{3\gamma} \cdot \ln x + e$? The honest answer is: we do not know the true functional form. If the specification is wrong, *everything collapses*. No amount of additional data can save us from a wrong model. Consider a concrete example: if the true data-generating process is logistic but we estimate a linear probability model, all parameter estimates are contaminated by functional form error.

Probit and Logit models impose specific link functions $P(y \mid X) = G(X'\beta)$. If the

true link function is something else, partial effects are inconsistent even asymptotically.

The core problem is the same in both cases: *any parametric model bets on a particular functional form.* If the bet is wrong, every inference built on it is wrong.

2.1.2 The Non-parametric Idea

The idea of non-parametric statistics is radical in its simplicity: let us *give up the functional form assumption entirely.* Instead of assuming $E[Y \mid X = x] = X'\beta$, we go back to the original question and estimate

$$g(x) \;=\; E[Y_i \mid X_i = x]$$

directly from data, without parametric restrictions.

This is not as exotic as it sounds. Recall from Chapter 1 that the potential outcome framework is *intrinsically non-parametric*: if we can directly estimate $E[Y_i \mid X_i = 1]$ and $E[Y_i \mid X_i = 0]$, we obtain a valid ATE estimate under unconfoundedness, with no functional form needed. Non-parametric methods generalize this idea to continuous X.

The tools introduced in this chapter (kernel regression, local polynomial regression, and series-based methods) provide systematic ways to estimate $g(x)$ non-parametrically. We will also introduce *semi-parametric models*, which occupy the middle ground: some components are parametric (and benefit from fast convergence rates), while others remain non-parametric (and absorb misspecification).

2.1.3 Notation

Throughout this chapter we adopt the following convention:

- x_i, y_i: **random variables**.
- X_i, Y_i: **realizations** (observed values); these are fixed in the sample and not random in our analysis.
- x, y: generic values or evaluation points.

An important implication: because X_i are realizations (not random in our context), integrals over random variables and sums over realizations do not interact: $\int x \sum_i^n X_i \, dx = \sum_i^n X_i \int x \, dx$.

2.2 Kernel-Based Methods

Kernel regression is the first non-parametric method we study, and it is beautifully intuitive. The central idea is to estimate the CEF *point by point*: instead of fitting one global function to all data, we estimate $E[Y_i \mid X_i = x]$ separately for each value of x, using only the observations whose X_i falls close to x.

We build up to this in three steps, each generalizing the previous one.

2.2.1 Step 1: Estimating the Cumulative Distribution Function

Before tackling kernel regression, let us start with a simpler problem: estimating a cumulative distribution function (CDF). Suppose we have an i.i.d. sample $\{X_1, \ldots, X_n\}$ of a scalar random variable X. The population CDF is

$$F(x) = P(X \leq x).$$

How do we estimate $F(x)$ at, say, $x = 3$? Go back to kindergarten! Just ask: what fraction of the data points are to the left of 3?

Looking at Figure 2.1, the answer is simply the proportion of dots to the left of the red vertical line. This gives us the *empirical CDF*:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x). \tag{2.1}$$

By the law of large numbers, $\hat{F}(x) \xrightarrow{p} F(x)$ for every x, and the Glivenko–Cantelli theorem guarantees uniform convergence over all x. No model needed, just counting.

2.2.2 Step 2: Estimating the Probability Density Function

Now consider the probability density function (PDF). Recall that the PDF is the derivative of the CDF:

$$f(x) = \frac{dF(x)}{dx} = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}.$$

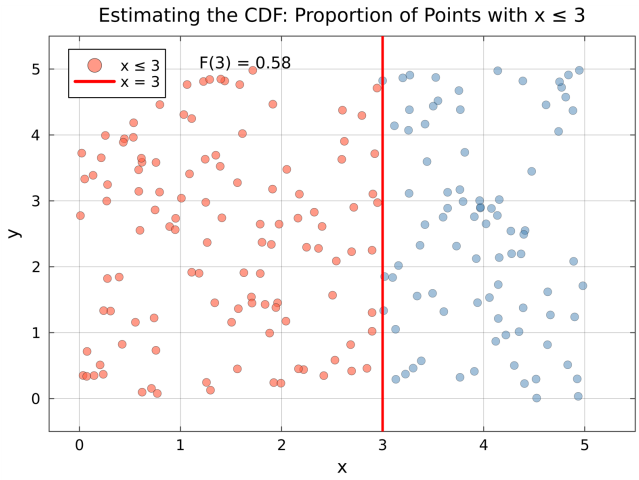


Figure 2.1: Estimating the CDF at $x = 3$. The red vertical line marks the evaluation point. The proportion of dots to the left of the line equals $\hat{F}(3)$.

The PDF represents the marginal increase in the CDF at a point: intuitively, how "crowded" the data are near x.

Replacing the population CDF with our empirical CDF gives a natural estimator:

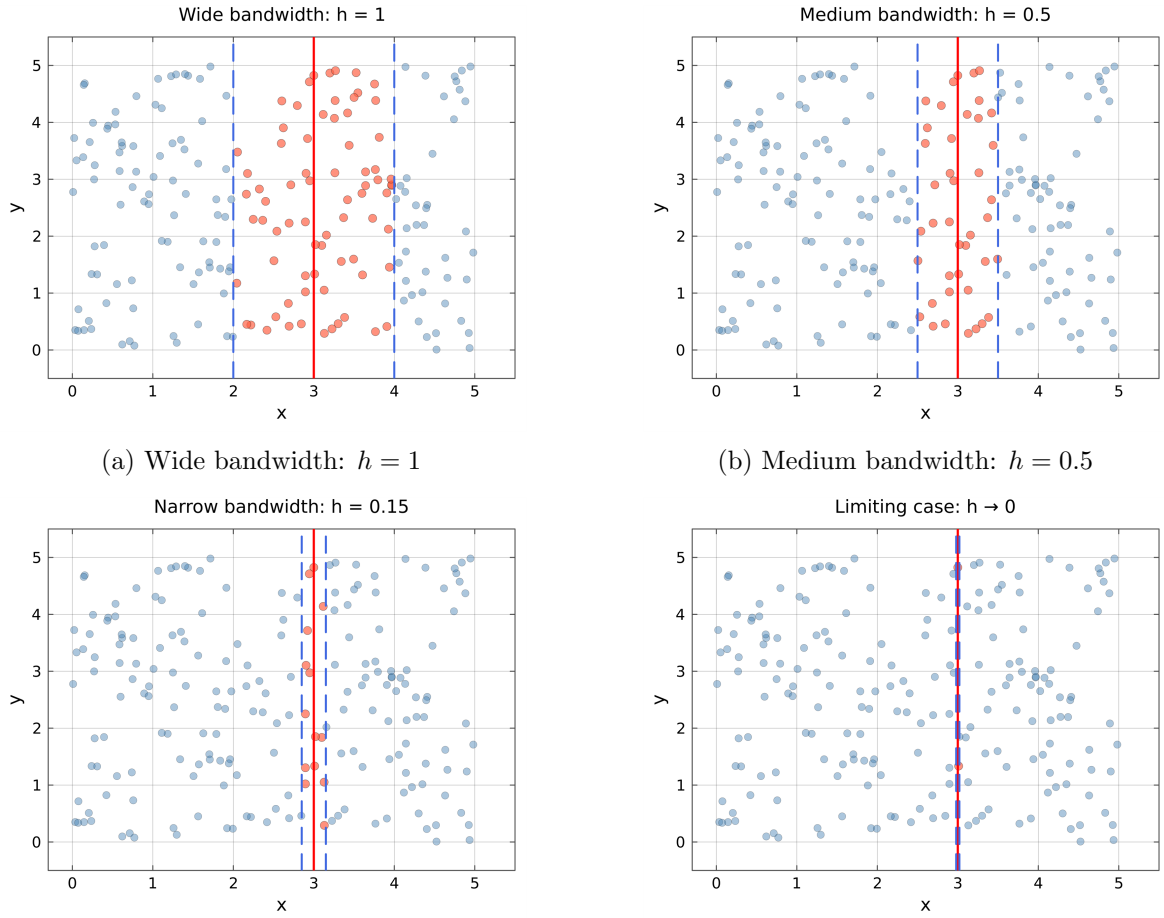
$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h}\,\mathbf{1}(x-h \le X_i \le x+h). \tag{2.2}$$

How to interpret this? We open a small window of width $2h$ centered at x, count the number of observations inside, divide by the window length $2h$ to get the density of observations per unit length, and divide by n to normalize. The parameter $h > 0$ is called the *bandwidth*.

The key insight is this: when n is large, we can choose a very small h to minimize the deviation from the evaluation point x and still have plenty of observations inside the window. This is what makes non-parametric estimation work.

Is it always good to have a smaller bandwidth? Not necessarily. A small bandwidth includes observations closer to the evaluation point, which reduces bias. However, it could significantly increase variance because the effective sample size shrinks. Specifically, with a very small bandwidth, there may be only a few observations and the estimation results will be totally driven by them. Therefore, with different draws of samples, the estimation result can be very volatile, creating a large variance. This is called the *bias-variance tradeoff*, which is a crucial concept in statistics. We will discuss this in more detail in the following sections and chapters.

Figure 2.2 illustrates how the bandwidth controls the window size. As h shrinks from 1 (top-left) toward zero (bottom-right), the window narrows and fewer observations contribute to the density estimate at $x = 3$. In the limit $h \to 0$, we recover the exact definition of the derivative.



(a) Wide bandwidth: $h = 1$

(b) Medium bandwidth: $h = 0.5$

(c) Narrow bandwidth: $h = 0.15$

(d) Limiting case: $h \to 0$

Figure 2.2: Effect of bandwidth on the estimation window (red line: $x = 3$; dashed blue lines: window boundaries $[x-h, x+h]$). A larger h includes more observations but introduces more bias; a smaller h reduces bias but increases variance.

2.2.2.1 The Uniform Kernel

Let us rewrite our density estimator in a more compact form. Define the *uniform kernel function*:

$$k(v) = \frac{1}{2}\,\mathbf{1}(|v| \leq 1).$$

Substituting $v = (X_i - x)/h$, the estimator (2.2) becomes

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k\left(\frac{X_i - x}{h} \right). \tag{2.3}$$

The function $k(\cdot)$ acts as a *weighting function*: it assigns equal weight to every observation inside the window and zero weight outside. This is the kernel density estimator (KDE) with a uniform kernel. The name "kernel" simply means "weight": *kernel is weight*.

2.2.2.2 General Kernel Functions

The uniform kernel treats all observations within the window equally, but this is not the only sensible choice. We might want to give *more* weight to observations very close to x and *less* weight to those near the edge of the window. This leads to general kernel functions.

A function $k : \mathbb{R} \to \mathbb{R}$ qualifies as a *kernel* if it satisfies two conditions:

1. $\int k(v)\, dv = 1$ (weights sum to one).
2. $k(v) = k(-v)$ (weights are symmetric around zero).

The first condition ensures the weights are normalized; the second ensures we treat observations to the left and right of x symmetrically. Common kernels include the following (Table 2.1):

Table 2.1: Common kernel functions used in non-parametric estimation.

Kernel	$k(v)$
Uniform	$\frac{1}{2}\,\mathbf{1}(\lvert v\rvert \le 1)$
Triangular	$(1 - \lvert v\rvert)\,\mathbf{1}(\lvert v\rvert \le 1)$
Epanechnikov	$\frac{3}{4}(1 - v^2)\,\mathbf{1}(\lvert v\rvert \le 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\, e^{-v^2/2}$

Notes: $\mathbf{1}(\cdot)$ denotes the indicator function. The first three kernels have compact support $[-1, 1]$; the Gaussian kernel has support $\mathbb{R}$ and assigns positive weight to all observations.

Figure 2.3 shows the shapes of these four kernels. The uniform kernel is flat inside the window. The triangular and Epanechnikov kernels give the highest weight to observations right at x and taper off toward zero at the window boundary. The Gaussian kernel is smooth everywhere and gives positive (but exponentially small) weight to all

observations, no matter how far from x.

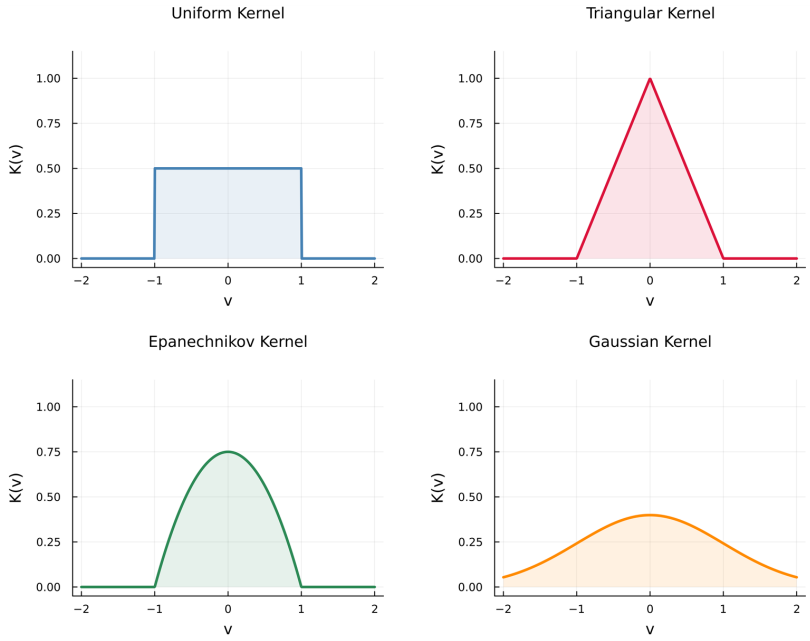


Figure 2.3: Four common kernel functions. Top left: uniform (constant weight within window). Top right: triangular (linear decay). Bottom left: Epanechnikov (parabolic; asymptotically optimal). Bottom right: Gaussian (unbounded support).

In practice, the Epanechnikov kernel is *optimal* in the sense of minimizing asymptotic mean squared error, and the triangular kernel is a close runner-up (it is the default in many regression discontinuity applications). But here is an important practical point:

Remark 2.2.1. The choice of kernel function has a relatively small effect on the final estimate. The bandwidth h matters far more. Choosing between Epanechnikov and Gaussian kernels is a second-order concern; choosing h well is first-order.

2.2.2.3 Multivariate Kernel Estimation

When X is a q-dimensional vector $X = (X^{(1)}, \ldots, X^{(q)})$, we extend the kernel to multiple dimensions using a *product kernel*:

$$K(v) = k(v_1)\, k(v_2) \cdots k(v_q), \quad v = (v_1, \ldots, v_q).$$

The multivariate kernel density estimator becomes

$$\hat{f}(x) = \frac{1}{n\, h_1 h_2 \cdots h_q} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where $h = (h_1, \ldots, h_q)$ is the vector of bandwidths. The product $h_1 h_2 \cdots h_q$ is the volume of the q-dimensional estimation window, a hyperrectangle.

In the two-dimensional case, the window is a rectangle with sides $2h_1 \times 2h_2$. Figure 2.4 illustrates: the red lines mark the evaluation point $(x_1, x_2) = (3, 3)$, the blue dashed lines mark the window boundaries, and only observations falling inside the rectangle receive nonzero weight.

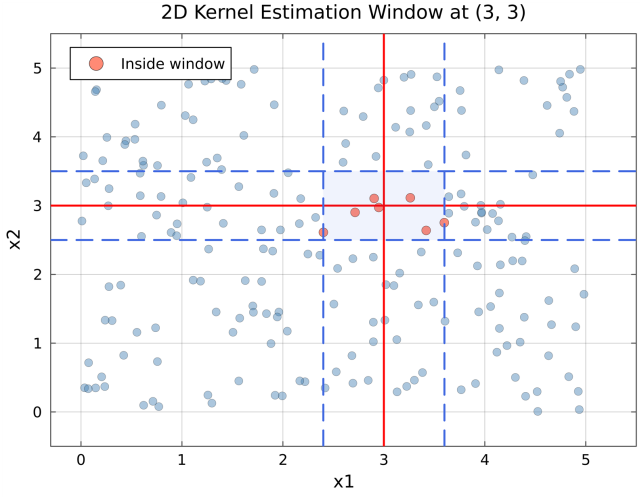


Figure 2.4: Two-dimensional kernel estimation window. The red lines mark the evaluation point $(x_1, x_2) = (3, 3)$; the blue dashed lines define the rectangular bandwidth window $[x_1 - h_1, x_1 + h_1] \times [x_2 - h_2, x_2 + h_2]$. Observations inside the rectangle receive positive kernel weight.

2.2.3 Step 3: Estimating the CEF with the Nadaraya–Watson Estimator

With CDF and PDF estimations in hand, we are ready for the main object of interest: the conditional expectation function $g(x) = E[Y \mid X = x]$.

The logic is the same as before: we estimate $g(x)$ *point by point*. For each evaluation point x, we form a weighted average of Y_i values using only observations whose X_i falls close to x. The kernel determines the weights; the bandwidth determines how "close"

counts.

Formally, recall that

$$g(x) = E[Y \mid X = x] = \frac{\int y\, f_{X,Y}(x, y)\, dy}{f_X(x)}.$$

Replacing each density by its kernel estimator and simplifying yields the *Nadaraya–Watson* (N-W) estimator (Nadaraya, 1964; Watson, 1964):

$$\hat{g}(x) = \sum_{i=1}^{n} Y_i\, \tilde{K}_h(X_i - x), \tag{2.4}$$

where the normalized kernel weight is

$$\tilde{K}_h(X_i - x) = \frac{K\left(\dfrac{X_i - x}{h}\right)}{\displaystyle\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}.$$

Nadaraya–Watson Estimator

The N-W estimator $\hat{g}(x)$ is a *locally weighted average* of the outcome Y_i for observations whose covariate X_i falls near x (within the bandwidth h). The weight given to observation i is proportional to how close X_i is to x, as measured by the kernel function. This is the non-parametric analogue of "computing a conditional mean."

Remark 2.2.2 (Homework). 1. Derive the N-W estimator formally from the kernel estimators of the joint and marginal densities. This is a bit involved; you may consult Hansen (2022) for guidance.

2. What does the N-W estimator reduce to when the uniform kernel is used? (Hint: it simplifies to a very familiar object.)

2.2.4 Asymptotic Theory for the N-W Estimator

Now that we have an estimator, we need to understand its statistical properties. Let $g(x) = E(Y \mid X = x)$ be the true CEF, $f(x)$ the marginal density of X at x, and $\sigma^2(x) = \mathrm{Var}(Y \mid X = x)$ the conditional variance.

Theorem 2.2.1 (Asymptotic Distribution of the N-W Estimator)**.** *Under standard regularity conditions, as* $n \to \infty$, $h_s \to 0$ *for* $s = 1, \ldots, q$, $nh_1 \cdots h_q \to \infty$, *and* $nh_1 \cdots h_q \sum_{s=1}^{q} h_s^6 \to 0$,

$$\sqrt{nh_1 \cdots h_q}\left(\hat{g}(x) - g(x) - \sum_{s=1}^{q} h_s^2 \, B_s(x)\right) \xrightarrow{d} N\left(0, \; \frac{\sigma^2(x)}{f(x)}\left(\int k(v)^2 \, dv\right)^q\right),$$

where the asymptotic bias component in direction s *is*

$$B_s(x) = \frac{\int v^2 k(v) \, dv}{2 \, f(x)}\left[2 \, \frac{\partial f(x)}{\partial x_s} \, \frac{\partial g(x)}{\partial x_s} + f(x) \, \frac{\partial^2 g(x)}{\partial x_s^2}\right].$$

This theorem looks complicated, but its key implications are surprisingly intuitive. Let us unpack them one by one.

2.2.4.1 The Bias-Variance Tradeoff

The asymptotic mean squared error (MSE) of $\hat{g}(x)$ has two components:

$$\mathrm{MSE}\big(\hat{g}(x)\big) \approx \underbrace{\left(\sum_{s=1}^{q} h_s^2 B_s(x)\right)^2}_{\text{Bias}^2} + \underbrace{\frac{1}{nh_1 \cdots h_q} \, \frac{\sigma^2(x)}{f(x)}\left(\int k(v)^2 \, dv\right)^q}_{\text{Variance}}.$$

This is the fundamental tension in all of statistics:

The Bias-Variance Tradeoff

Given data, it is *impossible* to have both lower bias and lower variance at the same time. The bandwidth h controls the tradeoff:

- **Increase** h (include more data points): variance falls because we average over more observations, but bias rises because we are averaging over a wider region where g may not be flat. This corresponds to Figure 2.2(a).
- **Decrease** h (use only nearby data points): bias falls because we are zooming in on a region where g is nearly constant, but variance rises because fewer observations contribute. This corresponds to Figure 2.2(c)–(d).

The same tradeoff applies to the choice of kernel: a more concentrated kernel reduces bias ($\int v^2 k(v) \, dv$ is smaller) but increases variance ($\int k(v)^2 \, dv$ is larger),

and vice versa.

This tradeoff is central not only to the choice of bandwidth and kernel function in kernel estimation, but also to model selection more broadly. A more flexible model, with a larger number of free parameters, may reduce bias at the cost of higher variance. By contrast, a more restrictive model tends to have lower variance but may suffer from greater bias. How should we determine the appropriate degrees of freedom for a model? We will return to this model selection issue in the next chapter.

2.2.4.2 The Curse of Dimensionality

As the dimension q increases, the volume of the window $h_1 \cdots h_q$ shrinks exponentially for any fixed $h_s < 1$. The variance term therefore *explodes* with q: we need exponentially more data to achieve the same precision. This is the *curse of dimensionality*. Non-parametric methods work well in low dimensions ($q = 1$ or 2), but become impractical for large q without additional structure.

2.2.4.3 Other Factors Driving Bias and Variance

Theorem 2.2.1 also reveals what features of the problem affect estimation quality:

1. **Slope and curvature of g.** Bias is large where $\partial g/\partial x_s$ (slope) or $\partial^2 g/\partial x_s^2$ (curvature) is large. Intuitively, if g is rapidly changing (first-order derivative), meanwhile the growth rate is also rapidly changing (second-order derivative), the values to the left and to the right of the evaluation point would not be "balanced." Averaging over a window of width h mixes together different values of g and introduces bias. For an extreme case, consider evaluating the function at a local maximum, denoted by x^*. On the one hand, the value of $g(x)$ varies rapidly around this point. On the other hand, the slope of $g(x)$ also changes sharply. Specifically, the derivative is positive to the left of x^* and turns negative immediately to the right. As a result, when we estimate $g(x^*)$, we rely on nearby observations from both sides whose function values are systematically much lower than $g(x^*)$ itself.
2. **Local density $f(x)$.** Both bias and variance are smaller where $f(x)$ is large; more local observations mean more information.
3. **Kernel concentration.** A more concentrated kernel (small $\int v^2 k(v)\, dv$) reduces bias but increases variance (large $\int k(v)^2\, dv$). This is another manifestation of the

bias-variance tradeoff.

2.2.5 Bandwidth Selection

The optimal bandwidth balances squared bias against variance. For the univariate case ($q = 1$), minimizing the asymptotic MSE gives an optimal bandwidth of order $h^* \sim n^{-1/5}$. Several practical methods exist:

- **Rule-of-thumb** (Silverman): assumes the density of X is Gaussian (a reference density) and gives a simple closed-form h.
- **Cross-validation**: choose h to minimize the leave-one-out prediction error

$$\text{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}_{-i}(X_i))^2,$$

  where $\hat{g}_{-i}$ omits observation i.
- **Plug-in methods**: estimate bias and variance components separately and then optimize.

By default, the Stata command `lpoly` uses a rule-of-thumb bandwidth (the estimator of Fan and Gijbels); cross-validation must be implemented manually.

How to choose the optimal number of terms K (in series methods) or the optimal bandwidth h is fundamentally the same question. We will revisit it in Chapter 3 when we discuss model selection and machine learning.

2.3 Local Polynomial Regression

The N-W estimator fits a *constant* locally: at each x, it simply computes the kernel-weighted average of Y_i. But what if the true $g(x)$ has a noticeable slope within the bandwidth window? Averaging over a slope introduces bias. A natural fix is to fit a *polynomial* locally rather than a constant.

The idea is intuitive: in linear regression, we fit one global linear function to all the data. In local polynomial regression, we fit a *piece-wise* polynomial function, interval by interval. For each evaluation point x, we choose only the observations very close to x (within the bandwidth), and fit a polynomial to just those observations.

For an evaluation point x, the local polynomial estimator of degree p solves:

$$\min_{b_0,b_1,\ldots,b_p} \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) \left(Y_i - b_0 - b_1(X_i - x) - b_2(X_i - x)^2 - \cdots - b_p(X_i - x)^p\right)^2. \tag{2.5}$$

The estimate of $g(x)$ is $\hat{g}(x) = \hat{b}_0$. Compared with global OLS, the only modification is the kernel weight $k\left(\frac{X_i - x}{h}\right)$: observations closer to x receive more weight, and observations far from x receive little or no weight.

- $p = 0$: local constant $\Rightarrow$ this is exactly the N-W estimator.
- $p = 1$: **local linear regression**, the most widely used version in practice.
- $p = 2$: local quadratic regression.

Figure 2.5 illustrates the idea. For a given evaluation point x (the vertical dotted line), only observations within the shaded orange region (the bandwidth window) contribute to the fit. A kernel (the bell curve at the bottom) assigns higher weight to observations near the center. A separate local polynomial is fitted at each point, and the estimate $\hat{g}(x)$ is the value of that polynomial evaluated at x. Tracing all these local estimates produces the full curve.

2.4 Series-Based Methods

Kernel-based methods (N-W and local polynomial) are flexible, but they have three practical drawbacks:

1. **Computational burden.** The estimator must be re-computed at every evaluation point; it is a point-by-point estimator, which can be slow for large datasets.
2. **Difficult to impose restrictions.** If economic theory implies that g is monotone or concave, it is hard to build that into a kernel estimator.
3. **Large sample requirement.** Non-parametric convergence rates are slow, so kernel methods can be imprecise with moderate sample sizes.

Series-based (or "sieve") methods address these problems by taking a different approach: instead of computing $g(x)$ point by point, they *approximate* $g(x)$ with a finite linear combination of known *basis functions*, then estimate the coefficients by OLS.

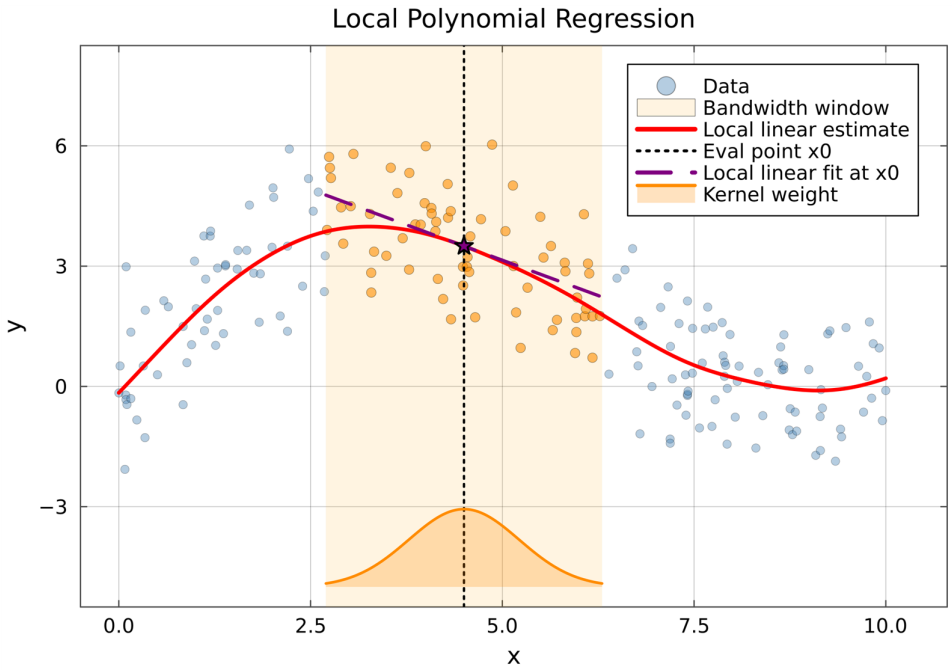


Figure 2.5: Local polynomial regression. For a given evaluation point x (vertical dotted line), the shaded orange region marks the bandwidth window. A kernel (the bell curve at the bottom) weights observations by proximity. The purple dashed segment is the local linear fit within the illustrated window; it is evaluated only at x (the star) to produce $\hat{g}(x)$. The red curve connects the local linear fits computed in the same way, with the same kernel and bandwidth, at every evaluation point.

2.4.1 Polynomial Basis

The simplest choice of basis comes from Taylor's theorem. Expanding the CEF at zero:

$$g(X) = \sum_{k=0}^{\infty} \frac{g^{(k)}(0)}{k!} X^k.$$

We approximate this infinite series by a *global polynomial* of degree K:

$$g(X) \approx \sum_{k=0}^{K} \theta_k \, p_k(X), \quad p_k(x) = x^k. \tag{2.6}$$

The vector $\{p_0, p_1, \ldots, p_K\}$ is the polynomial *basis*. This is just a standard regression model: we can estimate $\hat{\theta}_0, \ldots, \hat{\theta}_K$ by OLS and form $\hat{g}(x) = \sum_k \hat{\theta}_k p_k(x)$. Note the contrast: this is a *global* polynomial (one fit for all data), whereas local polynomial regression fits a separate polynomial at each x.

In the multivariate case with two regressors (X_1, X_2), the polynomial basis includes all cross terms: $\{1, X_1, X_2, X_1X_2, X_1^2, X_2^2, X_1X_2^2, X_1^2X_2, X_1^2X_2^2, \ldots\}$. The number of basis terms grows rapidly with dimension, reintroducing the curse of dimensionality.

2.4.2 Runge's Phenomenon

Polynomial series are the simplest basis, but they have a serious problem. Even for smooth, well-behaved functions, a high-degree polynomial can oscillate wildly near the boundary of the support. This is *Runge's phenomenon*.

Why does this happen? Power polynomials are "forced to vary somewhere": they are global functions that must accommodate the entire shape of g. When the polynomial order is high, the variation is pushed to the boundary, where it produces large oscillations.

Figure 2.6 shows this clearly: the red curve is the true function. The blue fifth-order polynomial fits the interior reasonably but starts to deviate near the boundary. The green ninth-order polynomial, despite having more terms, diverges *badly* at the endpoints.

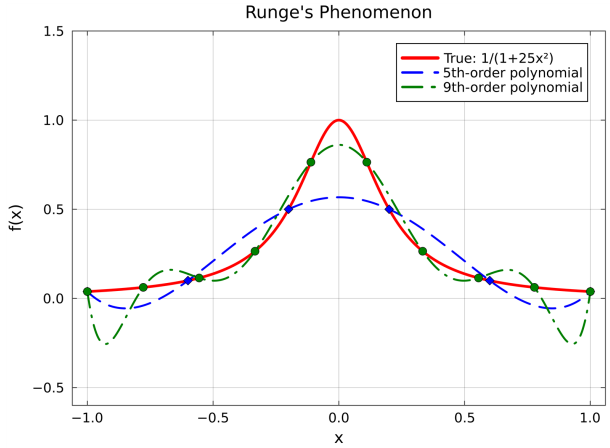


Figure 2.6: Runge's phenomenon. Red: true function. Blue: fifth-order polynomial approximation. Green: ninth-order polynomial approximation. Higher-order polynomials *worsen* the approximation near the boundaries despite fitting the interior better.

Runge's Phenomenon

High-degree polynomial series can approximate the interior of g well while producing large errors near the boundaries. The approximation error does *not* decline uniformly as K increases. In general, high-order polynomials behave very badly and should be avoided.

2.4.3 Fourier Basis

An alternative to polynomials is the Fourier (trigonometric) series. Any square-integrable function on a bounded interval can be represented as an infinite sum of sines and cosines; a finite truncation provides a good approximation. The Fourier basis is excellent for approximating periodic functions.

Figure 2.7 shows Fourier series with increasing numbers of terms: with only the constant term p_0, the fit is flat; adding two Fourier terms (p_1, p_2) captures the main shape; four terms $(p_1, \ldots, p_4)$ produce a much closer approximation.

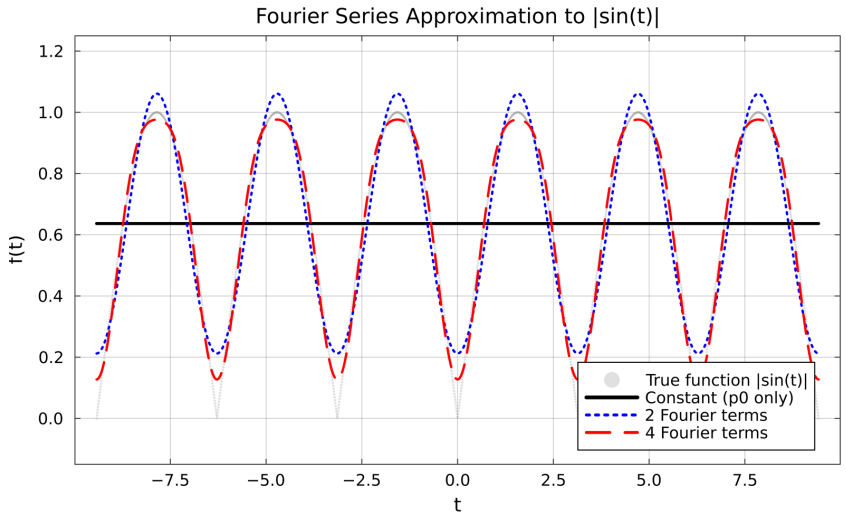


Figure 2.7: Fourier series approximation to a periodic function. Solid line: constant (p_0 only). Dotted line: two Fourier terms (p_1, p_2). Dashed line: four Fourier terms (p_1, p_2, p_3, p_4). More terms produce a better global approximation.

However, the Fourier basis has its own weakness: *Gibbs' phenomenon*. Near a discontinuity or sharp kink, the Fourier approximation exhibits persistent oscillations that do *not* disappear as more terms are added. Figure 2.8 shows this for a square wave: even with 50 Fourier terms, overshoot spikes of about 9% persist at each jump. The Fourier

basis is better than polynomials, but still not good at boundaries or jumping points.

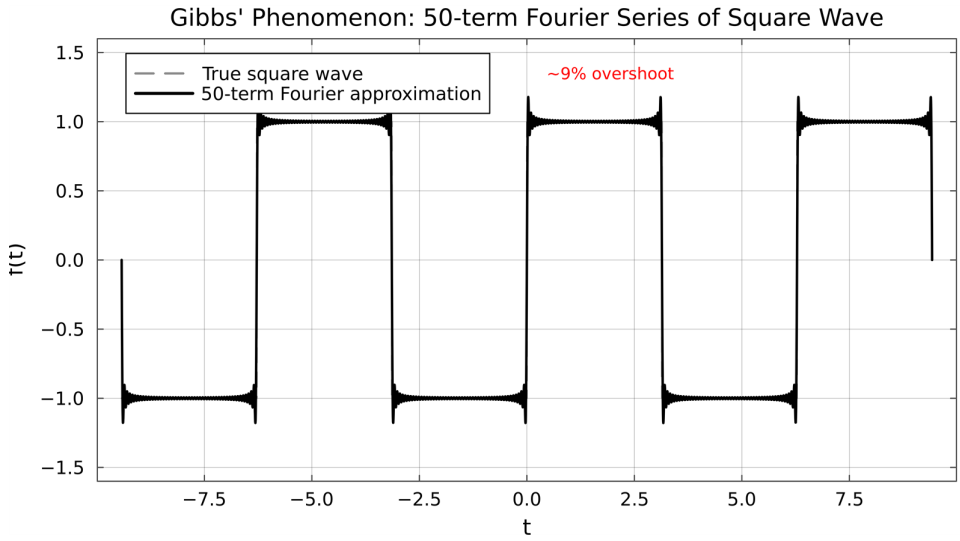


Figure 2.8: Gibbs' phenomenon: a 50-term Fourier series approximation to a square wave. Persistent oscillations near the jump discontinuities do not vanish as more terms are added.

2.4.4 Spline and Wavelet Bases

More sophisticated bases avoid both Runge's phenomenon and Gibbs' phenomenon:

- **Spline basis.** A spline is a piecewise polynomial joined smoothly at *knot* points. By controlling the placement of knots and the degree of smoothness at each knot, splines avoid the wild boundary behavior of global polynomials. Among series bases, spline bases are often a good default choice, especially for approximating smooth functions with local features.
- **Wavelet basis.** Wavelets are localized oscillations that capture both frequency and location. They are useful for functions that are smooth in some regions and rough in others.

These bases are rarely seen in mainstream applied econometrics, but are important tools in statistics and signal processing. For interested readers, Chen (2007) provides a thorough treatment.

2.4.5 Choosing the Number of Terms

How should one choose K, the number of basis terms? Too few terms underfit g (high bias); too many overfit (high variance). This is the same bias-variance tradeoff we

encountered with the bandwidth h.

In general, high-order polynomials behave very badly. Some other bases (splines, wavelets) are more forgiving. We will discuss this problem in detail in Chapter 3, which covers model selection and machine learning methods including cross-validation, AIC, BIC, and regularization.

2.5 Semi-parametric Models

Non-parametric methods are fully data-driven: they impose no prior structure on $g(x)$, no functional form assumptions, nothing. This generality comes at a cost: slow convergence rates, high variance, and large data requirements.

What if we want to impose *some* structure, but not the full parametric structure? This is the idea behind *semi-parametric models*: some components are specified parametrically, while others remain non-parametric. The parametric components converge at the fast $\sqrt{n}$ rate; the non-parametric components converge more slowly but absorb misspecification.

2.5.1 The Partially Linear Model

The most popular semi-parametric model is the *partially linear model* (PLM):

$$Y_i = X_i'\beta + g(Z_i) + u_i, \quad E[u_i \mid X_i, Z_i] = 0, \quad \text{Var}(u_i \mid X_i, Z_i) = \sigma^2. \tag{2.7}$$

Here, X_i enters the model *linearly* (with unknown coefficient β), and Z_i enters *non-parametrically* through the unknown function $g(\cdot)$. The researcher imposes linear structure where they are confident it is appropriate, and leaves the relationship with Z_i unrestricted.

2.5.2 Robinson's Two-Step Estimator

Robinson (1988) showed that β can be estimated at the parametric $\sqrt{n}$ rate despite the presence of the non-parametric component g. The key insight is a simple "partialling-out" argument.

Step 1: Remove the non-parametric component. Take the conditional expecta-

tion of (2.7) given Z_i:

$$E[Y_i \mid Z_i] = E[X_i \mid Z_i]'\beta + g(Z_i).$$

Subtract this from the original equation:

$$\underbrace{Y_i - E[Y_i \mid Z_i]}_{\equiv \tilde{Y}_i} = \underbrace{\big(X_i - E[X_i \mid Z_i]\big)'}_{\equiv \tilde{X}_i'} \beta + u_i.$$

The non-parametric component $g(Z_i)$ cancels out! What remains is a standard linear regression of $\tilde{Y}_i$ on $\tilde{X}_i$. In practice, $E[Y_i \mid Z_i]$ and $E[X_i \mid Z_i]$ are estimated non-parametrically (using kernel or local polynomial methods on Z_i), and then OLS is applied to the partialled-out equation.

Step 2: Recover the non-parametric component. Once $\hat{\beta}$ is obtained, estimate g by regressing the residual $Y_i - X_i'\hat{\beta}$ non-parametrically on Z_i:

$$Y_i - X_i'\hat{\beta} = g(Z_i) + u_i.$$

This can be estimated using any of the non-parametric methods introduced earlier in this chapter.

Robinson (1988)

In the partially linear model $Y_i = X_i'\beta + g(Z_i) + u_i$, Robinson's two-step estimator achieves the parametric $\sqrt{n}$ rate for $\hat{\beta}$, even though g converges at a slower non-parametric rate. The trick is to "partial out" the non-parametric component before estimating β.

2.5.3 A Common Pitfall: First-Stage Uncertainty

How do we conduct inference in this two-step procedure? Consider the variance of $\hat{g}(Z)$ in Step 2. Can we just use the standard errors from the non-parametric regression of $Y_i - X_i'\hat{\beta}$ on Z_i?

First-Stage Uncertainty is Not Free

No! The variable $Y_i - X_i'\hat{\beta}$ is *estimated*, not observed. It carries estimation error from Step 1. The standard errors from the Step 2 regression, which treat $Y_i - X_i'\hat{\beta}$

as known data, are therefore *too small*: they ignore the uncertainty propagated from the first step.
This mistake (treating a first-stage estimate as a known quantity in the second stage) is one of the most common errors in empirical work and leads to invalid inference. Watch out for the standard error estimation whenever you have a multi-step procedure!

For $\hat{\beta}$, Robinson (1988) derived the correct asymptotic variance that accounts for first-stage estimation error. For $\hat{g}(z)$ in Step 2, usually, no simple closed-form variance formula is available. This is because the residual $Y_i - X_i'\hat{\beta}$ is a combination of data and estimated quantities, making analytical variance derivation intractable.

There are many cases when closed-form variance is not available at all. When this happens, we need a different approach for inference: the *bootstrap*.

2.6 Bootstrap

2.6.1 Why Bootstrap?

The bootstrap is a non-parametric method for inference. It is used precisely when closed-form standard errors are unavailable or unreliable, which, as we just saw, is a common situation in semi-parametric and non-parametric models.

The idea is elegant: instead of deriving the sampling distribution of our estimator analytically, we *simulate* it. We use the sample itself as a stand-in for the population, and by repeatedly resampling from the observed data, we approximate the sampling distribution of any statistic.

2.6.2 The Bootstrap Algorithm

1. **Draw bootstrap samples.** Given the observed sample $\{(Y_i, X_i)\}_{i=1}^n$ of size n, draw R new samples of size n *with replacement*. Index each bootstrap sample by $r = 1, \ldots, R$.
2. **Compute the statistic.** For each bootstrap sample r, compute the estimate of interest, $\hat{g}^{(r)}(x)$.

3. **Estimate the variance.** The *bootstrapped variance* is:

$$\hat{V}(x) = \frac{1}{R-1}\sum_{r=1}^{R}\left[\hat{g}^{(r)}(x) - \bar{g}(x)\right]^2, \qquad \bar{g}(x) = \frac{1}{R}\sum_{r=1}^{R}\hat{g}^{(r)}(x).$$

2.6.3 Bootstrap Confidence Intervals

Normal-approximation interval. Use $\hat{V}(x)$ as the variance estimate and form a $100(1-\alpha)\%$ interval as $\hat{g}(x) \pm z_{\alpha/2}\sqrt{\hat{V}(x)}$. However, this approach relies on asymptotic normality, which may not be accurate in finite samples.

Percentile interval (preferred). A better approach: stack the R bootstrap estimates $\{\hat{g}^{(1)}(x), \ldots, \hat{g}^{(R)}(x)\}$ to form an empirical distribution. Let q^*_α denote the α-quantile of this distribution. The bootstrap percentile confidence interval is:

$$\left[q^*_{\alpha/2},\ q^*_{1-\alpha/2}\right].$$

Bootstrap Percentile Interval

The bootstrap percentile interval $[q^*_{\alpha/2}, q^*_{1-\alpha/2}]$ uses the empirical quantiles of the bootstrap distribution directly. It does not assume normality and is more robust in practice. This is the preferred method for constructing confidence intervals in non-parametric settings.

2.7 Applications and Summary

Where should we apply non-parametric methods? We offer a summary below:

- Anything related to estimation of the CEF $g(x) = E[Y \mid X = x]$.
- The potential outcome framework is intrinsically non-parametric: we can estimate ATE/ATT without regression.
- Non-parametric inference in complicated models (via bootstrap).
- When prediction and fit matter more than structural interpretation: stock price prediction, machine learning, RDD fitting.

We will encounter non-parametric tools throughout this course. For example:

- **Matching estimators (Chapter 1):** kernel smoothing is used to estimate propen-

sity scores and conditional means.

- **Machine learning (Chapter 3):** series methods and cross-validation underlie regularized regression and tree-based methods.
- **Regression discontinuity (Chapter 10):** local polynomial regression is used to estimate the treatment effect at the cutoff.

However, non-parametric methods have important limitations. As we have discussed: *it is not always better to make the model more flexible.*

- **A correctly specified parametric model is more efficient.** If the true CEF is linear, OLS beats any non-parametric estimator.
- **Hard to incorporate restrictions.** Economic theory often implies shape restrictions (monotonicity, concavity) that are difficult to enforce non-parametrically.
- **Large samples required.** With small n, non-parametric estimates can be noisy unless the bandwidth is large, which reintroduces bias.

Homework Problems

1. **Nadaraya–Watson Derivation.** Derive the N-W estimator from first principles. Specifically, write the joint density of (X, Y) as $f_{X,Y}(x, y)$, express $g(x)$ as a ratio of integrals involving $f_{X,Y}$ and f_X, replace each density with its kernel estimator, and simplify to obtain equation (2.4).

2. **Uniform Kernel N-W.** Show that the N-W estimator with a uniform kernel reduces to the simple sample mean of Y_i values in the window $[x - h, x + h]$:

$$\hat{g}(x)\big|_{\text{uniform}} = \frac{\sum_{i=1}^{n} Y_i \, \mathbf{1}(|X_i - x| \leq h)}{\sum_{i=1}^{n} \mathbf{1}(|X_i - x| \leq h)}.$$

Practice Example: Non-parametric Regression in Stata

This practice example illustrates the use of non-parametric methods in Stata. We simulate a dataset with a known non-linear CEF, show that global polynomial OLS systematically misses the non-linearity, and then apply Stata's `lpoly` command (local polynomial regression) to recover the true CEF. We compare the in-sample RMSE

of global quadratic OLS, global quartic OLS, and local linear regression against the irreducible noise floor.

Setup

The data-generating process is:

$$Y_i = g(X_i) + \varepsilon_i, \quad g(x) = 3\sin(2x) + 0.5\,x^2 - 2,$$
$$\varepsilon_i \sim N(0,1), \quad X_i \sim \text{Uniform}[0,5], \quad n = 500.$$

The true CEF is the non-linear function $g(x) = 3\sin(2x) + 0.5x^2 - 2$, which combines a sinusoidal oscillation with a quadratic trend. No global polynomial of low degree can capture this shape well.

Learning Objectives

1. A simple OLS regression of Y on X and X^2 provides a rough fit but fails to capture the oscillatory component of g.
2. Local polynomial regression (`lpoly` with `degree(1)` and `degree(3)`) recovers the true CEF closely.
3. The bias-variance tradeoff is visible: a very narrow bandwidth ($h = 0.03$) tracks individual data points but is extremely noisy; a wide bandwidth ($h = 2.0$) is smoother but misses local features.
4. In-sample RMSE comparison: global quadratic OLS, global quartic OLS, and local linear regression are compared against the irreducible noise

   $$\text{RMSE}_{\text{true}} = \sqrt{n^{-1}\textstyle\sum_{i=1}^{n}(Y_i - g(X_i))^2} \approx \sigma = 1.$$

   Local linear regression achieves RMSE closest to the noise floor.

Replication Code

The Stata do-file `src/chapter 2/ch2_nonpar.do` reproduces all results. The simulated dataset is saved to `data/chapter 2/ch2_nonpar.dta`.

Chapter 3

Machine Learning and Model Selection

3.1 Introduction

In Chapter 2 we assembled a powerful toolkit of non-parametric and semi-parametric methods: kernel regression, local polynomial regression, series estimation, and the partially linear model. Together with the linear regression from Chapter 1, we now have many tools in our toolbox. A natural and important question follows: *which tool should we choose?*

This is a model selection question, and it arises even when we restrict ourselves to a single tool. Take ordinary least squares (OLS) as an example. A linear specification says $y = x\beta+\varepsilon$. Why linear? Is it because linearity is simple? Why not $y = \ln x+x^3+\varepsilon$? Consider the Mincer wage equation, in which we regress *wage* on *edu*, *exp*, and exp^2. Why include the square of experience but not, say, edu^3? The honest answer is that most of the time we are making these choices by habit or convention rather than by any systematic criterion. The functional form is our assumption, and, as we saw in Chapter 2, if the assumption is wrong, everything collapses.

For a long time, model selection was largely ignored in applied economics. The main reason was data availability: when your dataset had a few hundred observations and a handful of variables, it was not feasible for you to further split your small dataset into a training (estimation) set and a validation set. Thus, the scope for model selection

was limited. Nowadays, the world has changed. More and more datasets are available with large sample sizes and high-dimensional covariate spaces. Big data brings more chances, but also more decisions to make. We should take model selection seriously.

There are broadly two approaches to choosing a model. The first is a **data-driven approach**: we use the data themselves to select the model, without injecting prior knowledge about causal relationships. This is the machine learning philosophy. The second is a **structure-based approach**: we use economic theory, causal logic, or a Directed Acyclic Graph (DAG) to determine which variables belong in the model. In this chapter we focus on the first approach. We select models using only data, without putting our economic knowledge or theory into the process. The structure-based approach will be the topic of Chapter 4.

Before introducing any specific machine learning algorithm, we need to understand a fundamental statistical concept that governs all of model selection: the *bias-variance tradeoff*. We encountered this idea briefly in Chapter 2 when we discussed the choice of bandwidth in kernel regression. Here we develop it more generally and show that it is the central principle behind every model selection method.

3.2 The Bias-Variance Tradeoff

3.2.1 Why Not Always Use the Most Complicated Model?

Consider three models of increasing flexibility:

$$\text{A traditional linear model:} \qquad y = x\beta + \varepsilon, \tag{3.1}$$

$$\text{A model with a quadratic term:} \qquad y = x\beta + x^2\alpha + \varepsilon, \tag{3.2}$$

$$\text{A non-parametric model:} \qquad y = g(x) + \varepsilon. \tag{3.3}$$

Why not always use the second or the third one? More flexibility means less risk of misspecification, so it seems like a free lunch.

To sharpen the question, consider two nested linear models:

$$\text{Model A:} \qquad y = x_1'\beta + \varepsilon, \tag{3.4}$$

$$\text{Model B:} \qquad y = x_1'\beta_1 + x_2'\beta_2 + \varepsilon. \tag{3.5}$$

Model B includes everything in Model A plus additional regressors x_2. Is it always better to have the more complicated model?

The answer is *no*. The reason is that more flexibility helps with one problem but hurts with another. Adding x_2 reduces the risk of misspecification; this lowers *bias*. But it also makes the estimated model more sensitive to the particular sample at hand; this raises *variance*. This tension between bias and variance is the central theme of this chapter.

3.2.2 Decomposing Prediction Error

To make the tradeoff precise, assume the true data-generating process is

$$Y = f(X) + \varepsilon, \qquad E[\varepsilon \mid X] = 0, \qquad \text{Var}(\varepsilon \mid X) = \sigma_\varepsilon^2.$$

We train a model $\hat{f}(x)$ on some data. Because the training data are random, $\hat{f}(x)$ is itself a random object. If we drew a different sample, we would get a different fitted model: $\hat{f}^1(x)$, $\hat{f}^2(x)$, and so on. The expectation $E[\hat{f}(x)]$ is taken over these different samples; it is the average model we would get if we could repeat the estimation many times.

How good is the model at predicting a new observation? The standard measure is the *prediction mean squared error* (PMSE) at a point x_0:

$$E\Big[(Y - \hat{f}(x_0))^2 \mid X = x_0\Big].$$

This measures how far, on average, our prediction $\hat{f}(x_0)$ is from the actual outcome Y. The following decomposition is the cornerstone of model selection theory.

Bias-Variance Decomposition

$$E\Big[(Y - \hat{f}(x_0))^2 \mid X = x_0\Big] = \underbrace{\sigma_\varepsilon^2}_{\text{irreducible error}} + \underbrace{\big[E\hat{f}(x_0) - f(x_0)\big]^2}_{\text{Bias}^2} + \underbrace{E\big[\hat{f}(x_0) - E\hat{f}(x_0)\big]^2}_{\text{Variance}}. \tag{3.6}$$

Prediction error = Irreducible error + Bias^2+ Variance.

Let us understand each component.

Irreducible error (σ_ε^2). This is the noise inherent in the data. No matter how clever our model is, we can never predict better than σ_ε^2 because the outcome Y is genuinely random around $f(X)$. This term sets the floor for prediction error.

Bias $[E\hat{f}(x_0) - f(x_0)]$. Bias measures the systematic error: how far, on average, our model's prediction is from the truth. A simple model like linear OLS applied to a non-linear DGP will systematically miss the curvature, producing large bias. A very flexible model can track the curvature closely, producing small bias.

Variance $E[\hat{f}(x_0) - E\hat{f}(x_0)]^2$. Variance measures how much the fitted model changes from sample to sample. A simple model like linear OLS is stable: it produces similar fitted lines regardless of the specific sample drawn. A very flexible model (say, a twentieth-order polynomial) is highly sensitive to the particular data points: adding or removing a few observations can change the fitted curve dramatically.

The fundamental tension is now clear:

- Increasing model complexity $\Rightarrow$ Bias $\downarrow$ but Variance $\uparrow$.
- A super complicated model $\Rightarrow$ Variance $\uparrow\uparrow\uparrow$. The model becomes so sensitive to the training data that it changes wildly from one sample to the next.
- Such a model *overfits* the current data: it achieves excellent in-sample fit but poor out-of-sample prediction.

Good model selection means finding the sweet spot that minimizes the *total* prediction error: the sum of bias squared and variance.

3.2.3 An Example of Overfitting

Let us make the discussion concrete with a simple simulation. Consider the data-generating process

$$Y = 1 + 1.5X + \varepsilon, \qquad \varepsilon \sim N(0, 3600).$$

This is a linear relationship, but a very noisy one: the standard deviation of the noise is 60, which is large relative to the signal. Suppose we draw 30 observations from this process and try to fit polynomials of increasing order.

With a **first-order (linear)** polynomial, the fitted line lies close to the true DGP. There is some discrepancy (the line does not pass through every data point), but this is appropriate because the data points contain noise. The fit captures the signal and ignores the noise.

With a **second-order (quadratic)** polynomial, the fitted curve bends slightly, picking up a small amount of curvature that exists in the noise but not in the true DGP. The improvement in R^2 is marginal.

By the **fifth-** and **sixth-order**, the fitted curve is clearly oscillating. It tries to pass more data points by twisting the curve, but the wiggles have nothing to do with the true DGP; they are artifacts of the noise.

Figure 3.1 illustrates this progression. In the first-order panel the fitted line (red) nearly coincides with the true DGP (gray dashed). As the polynomial order rises to 5 and 6, the fitted curve begins to deviate from the straight line, bending to chase individual data points.

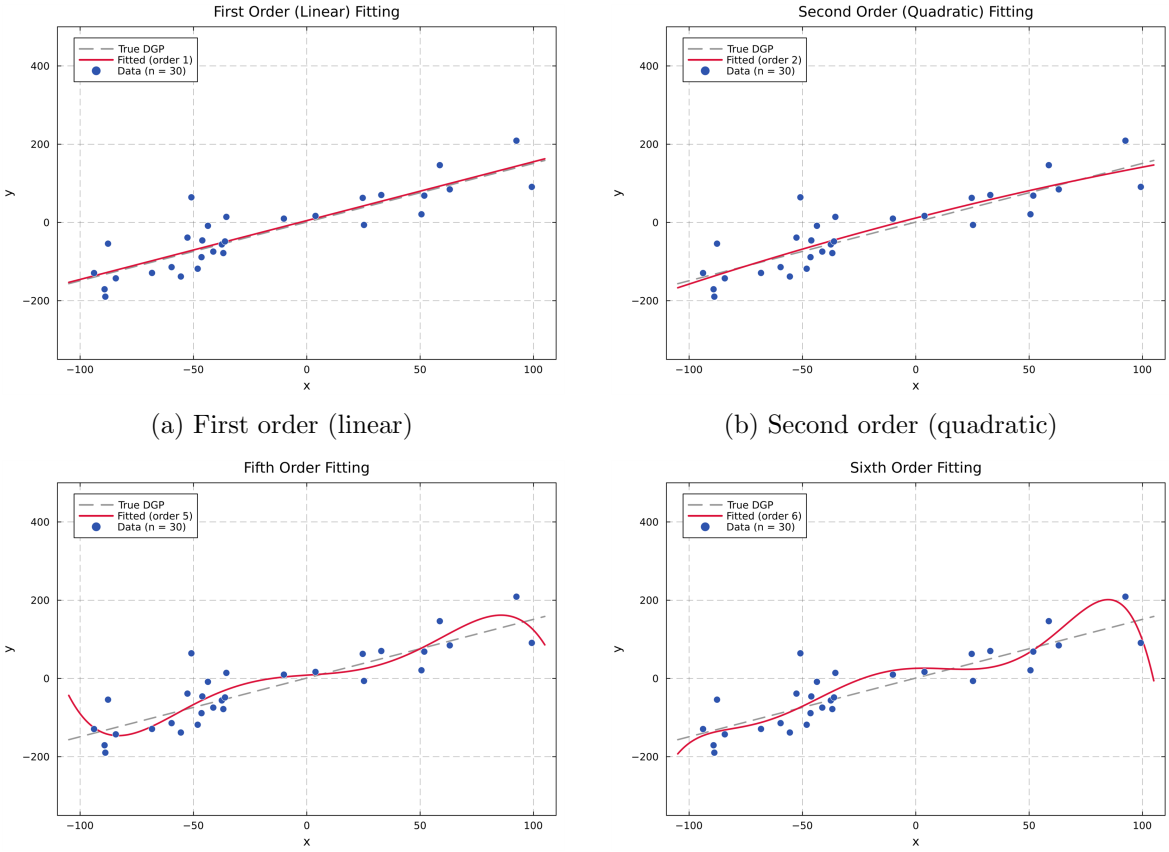
Overfitting

High-order polynomials pick up **noise, not signals**. They achieve excellent in-sample fit but terrible out-of-sample prediction. This is the defining symptom of overfitting. The model's bed has been carved into the exact shape of one night's sleeper at a specific time point: it fits that sleeper perfectly at that moment, but anyone who moves will find the bed unusable.

This example echoes our experience from Chapter 2. Recall Runge's phenomenon: a high-degree global polynomial oscillates violently at the boundary, even though the function it tries to approximate is smooth. And the Gibbs phenomenon: a Fourier series overshoots at a discontinuity no matter how many terms we add. Both are instances of the same underlying principle: *more parameters do not always improve approximation*. The extra flexibility can be wasted on fitting noise rather than signal.

3.3 Goodness of Fit and Cross-Validation

The overfitting example teaches us that we cannot simply maximize in-sample fit. We need a criterion that rewards good fit but penalizes unnecessary complexity. In this section we review several such criteria.



(a) First order (linear)

(b) Second order (quadratic)

(c) Fifth order

(d) Sixth order

Figure 3.1: Polynomial fitting of orders 1, 2, 5 and 6. The true DGP is $Y = 1 + 1.5X + \varepsilon$ with $\sigma = 60$ and $n = 30$. Gray dashed: true relationship. Red solid: fitted polynomial. As the order increases, the fitted curve begins to deviate from the straight-line truth, tracking noise rather than signal.

Figure 3.2: The bias-variance tradeoff and overfitting. As model complexity increases, training error decreases monotonically, but test error eventually rises, indicating overfitting. The optimal model minimizes total prediction error (test error).

3.3.1 Traditional Information Criteria

The most familiar approach is to modify the in-sample fit measure with a penalty for the number of parameters.

Adjusted R^2. Recall from introductory econometrics that R^2 never decreases when we add a regressor, even an irrelevant one. This is because OLS can always use an additional coefficient to reduce the residual sum of squares, even if only by fitting noise. The adjusted R^2 corrects for this by dividing both the residual and total sums of squares by their respective degrees of freedom. It can decrease when a useless regressor is added, because the penalty for losing a degree of freedom exceeds the gain in fit. This is the simplest form of complexity penalization.

AIC (Akaike Information Criterion). The AIC provides a more principled penalty:

$$\text{AIC} = 2k + n \ln(\text{RSS}/n),$$

where k is the number of estimated parameters and RSS is the residual sum of squares. The term $n \ln(\text{RSS}/n)$ measures goodness of fit (smaller RSS means better fit), while $2k$ penalizes complexity. We select the model with the smallest AIC. The factor of 2 in the penalty term reflects an asymptotic correction for the optimism of the maximized

log-likelihood as an estimator of the expected Kullback–Leibler divergence from the true model.

BIC (Bayesian Information Criterion). The BIC is similar in spirit but uses a stronger penalty:

$$\text{BIC} = k \ln(n) + n \ln(\text{RSS}/n).$$

The penalty $k \ln(n)$ grows with the sample size, so BIC penalizes additional parameters more heavily in large samples. BIC is motivated by the Bayesian approach to model comparison. In practice, BIC tends to select simpler models than AIC, which is desirable when the true model is sparse.

3.3.2 Cross-Validation

AIC and BIC penalize complexity analytically. Cross-validation (CV) takes a different approach: it directly measures out-of-sample prediction performance by splitting the data into pieces.

The idea is beautifully simple. We want to know how well a model predicts on data it has never seen. So we deliberately set aside some observations, train the model on the rest, and then check how well the model predicts the held-out observations. If the model has overfit the training data, it will perform poorly on the held-out data, because the noise patterns it memorized are different in the new observations.

The most common implementation is **K-fold cross-validation**:

1. Divide the sample randomly into K equal-sized groups, called "folds."
2. Pick fold k (one fold in K) as the *validation sample*. Use the remaining $K-1$ folds as the *training sample* to estimate the model.
3. Compute the mean squared prediction error MSE_k on fold k, that is, the average squared difference between the actual Y values in fold k and the model's predictions for those observations.
4. Repeat for each fold $k = 1, 2, \ldots, K$, so that every fold serves as the validation sample exactly once.
5. Average the K error measures:

$$\text{CV} = \frac{1}{K} \sum_{k=1}^{K} \text{MSE}_k.$$

Common choices are $K = 5$ or $K = 10$. The extreme case $K = n$ (each observation is its own fold) is called *leave-one-out* cross-validation.

Cross-Validation

CV measures the quality of *out-of-sample* prediction. We use data that were not involved in the estimation to evaluate the model. A smaller CV value means a model that generalizes better to new data. CV is particularly valuable because it directly targets what we care about (prediction on unseen observations) rather than relying on asymptotic approximations.

3.3.3 From Goodness of Fit to Machine Learning

We now have several standards for what constitutes a "good" model: adjusted R^2, AIC, BIC, and cross-validation. These criteria tell us *how to evaluate* a model, but they do not tell us *how to search* for one. If we have 20 potential regressors, there are $2^{20} \approx 1$ million possible subsets to consider. Evaluating all of them is computationally expensive and statistically problematic.

This is where machine learning enters the picture. Machine learning algorithms provide *automatic, efficient search procedures* for finding good models. They take the goodness-of-fit criteria we have just discussed and embed them into algorithms that explore the model space intelligently.

3.3.4 What Is Machine Learning?

What is machine learning? For an economist's perspective on the question, see Athey and Imbens (2019); the Wikipedia definition is instructive:

> "Machine learning (ML) is an umbrella term for solving problems for which development of algorithms by human programmers would be cost-prohibitive, and instead *the problems are solved by helping machines 'discover' their own 'algorithms'*, without needing to be explicitly told what to do by any human-developed algorithms."

In economics, the use of machine learning is more specific. The main question is: *how complicated should the model be?* Given a set of potential predictors X, how should we predict Y? When Y is discrete (e.g., default vs. no default), this is a **classification**

problem. When Y is continuous (e.g., house prices), this is a **prediction (regression)** problem.

There are many machine learning algorithms. We will introduce three families that are most relevant for economists: *penalized regression*, *tree-based methods*, and *neural networks*.

3.4 Penalized Regression

3.4.1 The Problem of Too Many Regressors

Let us start with a setting that is close to what economists already know: linear regression. Consider the model $y_i = x_i'\beta + \varepsilon_i$. What happens when the number of potential regressors is very large?

This is not a hypothetical concern. Imagine you have a household survey with 1,000 questions. Each answer is a potential regressor. Which ones should you include in your wage equation? OLS estimates

$$\hat{\beta}^{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2.$$

Because OLS minimizes the sum of squared residuals (SSR), every additional regressor helps: more β's mean more flexibility, which means a smaller SSR. But we know that a smaller SSR does not mean a better model; it may just mean more overfitting. We need a mechanism to *penalize* the use of additional regressors.

3.4.2 The Penalized Objective Function

The key idea is to modify the OLS objective by adding a *penalty term* that discourages large or numerous coefficients:

Penalized Regression

$$\hat{\beta}^{\text{Pen}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda \left(\|\beta\|_p\right)^p, \tag{3.7}$$

where $\|\beta\|_p = \left(\sum_j |\beta_j|^p\right)^{1/p}$ is the ℓ_p-norm, $(\|\beta\|_p)^p = \sum_j |\beta_j|^p$, and $\lambda \geq 0$ is the *tuning parameter*.

The objective function now has two parts. The first part, $\sum_i (y_i - x_i'\beta)^2$, is the usual SSR; it wants β to be large enough to fit the data well. The second part, $\lambda\|\beta\|_p^p$, is the penalty; it wants β to be small. The estimator must balance these two competing goals. The tuning parameter λ controls the tradeoff: larger λ means more shrinkage (simpler model, lower variance, higher bias); $\lambda = 0$ recovers OLS.

The two most important special cases correspond to $p = 1$ and $p = 2$.

3.4.2.1 LASSO Regression ($p = 1$)

When $p = 1$, the penalty is $\lambda \sum_j |\beta_j|$. This is called the **Least Absolute Shrinkage and Selection Operator** (LASSO), introduced by Tibshirani (1996).

The ℓ_1 penalty has a remarkable property: it can shrink some coefficients *exactly to zero*. This means that LASSO does not just shrink coefficients; it performs *variable selection*. Regressors with small predictive power are dropped entirely. In our 1,000-question survey example, LASSO might select, say, 15 variables and set the remaining 985 coefficients to exactly zero.

Why does ℓ_1 produce exact zeros while ℓ_2 does not? The geometric intuition is instructive. The ℓ_1 constraint set (the "diamond" $\sum |\beta_j| \leq c$) has sharp corners along the axes. The OLS contour is an ellipsoid. When the ellipsoid first touches the diamond, it is very likely to touch at a corner, where one or more coordinates are exactly zero. The ℓ_2 constraint set (the "ball" $\sum \beta_j^2 \leq c$) is smooth, so tangency almost surely occurs at a point of the smooth boundary away from the axes, where no coordinate is exactly zero.

3.4.2.2 Ridge Regression ($p = 2$)

When $p = 2$, the penalty is $\lambda \sum_j \beta_j^2$. This is called **Ridge regression**.

Ridge regression shrinks all coefficients toward zero but never sets any exactly to zero. In our survey example, Ridge would keep all 1,000 regressors but give small coefficients to the irrelevant ones. Ridge is particularly useful when regressors are highly correlated: in such settings, OLS estimates are unstable (small changes in the data produce large

swings in $\hat{\beta}$), while Ridge stabilizes them through shrinkage.

3.4.2.3 Choosing λ

How do we choose the tuning parameter λ? Cross-validation. We try a grid of λ values, compute the CV score for each, and pick the λ that minimizes CV. This directly targets out-of-sample prediction quality, which is exactly what we want.

3.4.3 The Elastic Net

In practice, a popular choice is the **Elastic Net**, which combines the ℓ_1 and ℓ_2 penalties:

$$\hat{\beta}^{\text{EN}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \left[\alpha \|\beta\|_1 + (1-\alpha) \left(\|\beta\|_2 \right)^2 \right], \tag{3.8}$$

where $\alpha \in [0, 1]$ controls the mix between LASSO ($\alpha = 1$) and Ridge ($\alpha = 0$). The Elastic Net inherits the variable selection property of LASSO while retaining the stability of Ridge when regressors are correlated.

3.5 Tree-Based Methods

3.5.1 The Idea: Partition and Average

Penalized regression stays within the linear model and controls complexity by shrinking coefficients. Tree-based methods take a completely different approach. The idea is disarmingly simple:

1. Partition the covariate space X into a set of rectangular regions.
2. Within each region, predict Y by the average of the Y values in that region.

This is a **Classification and Regression Tree** (CART; see Chapter 9 of Hastie et al., 2009). The name reflects the fact that the partition can be represented as a binary tree: each internal node is a yes/no question about one covariate ("Is $X_1 \leq 3.5$?"), and each terminal node (leaf) contains a predicted value.

Formally, we partition the space into M disjoint regions $R_1, \ldots, R_M$ and define the

predicted value within each region as the sample mean:

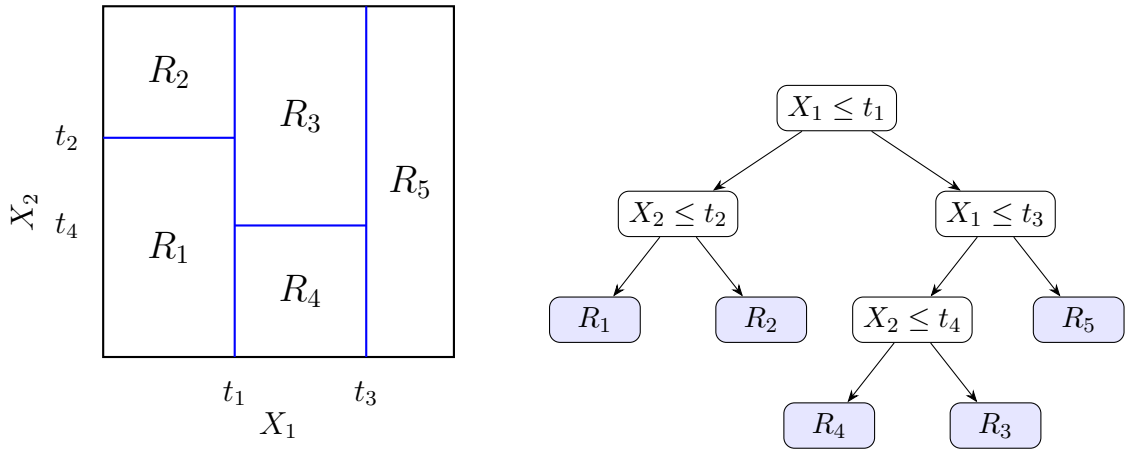
$$\hat{f}(x) = \sum_{m=1}^{M} \hat{c}_m \, \mathbf{1}(x \in R_m), \qquad \hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i,$$

where $N_m = |\{x_i \in R_m\}|$ is the number of observations in region m.

3.5.2 Recursive Binary Partitions

How do we find a good partition? The answer is **recursive binary partitions**. At each step we split one existing region into two halves by choosing a variable and a cut point. The process produces a binary tree: each internal node is a yes/no question about one covariate, and each terminal node (leaf) contains a predicted value.

Figure 3.3 illustrates a concrete example with two covariates X_1 and X_2. The first split is at $X_1 = t_1$; the left region is then split at $X_2 = t_2$; and so on. The resulting five regions $R_1, \ldots, R_5$ correspond exactly to the five leaves of the binary tree shown alongside.



(a) Partition of the (X_1, X_2) space

(b) Corresponding binary tree

Figure 3.3: Recursive binary partitioning. Each split divides one region into two by choosing a variable and a cut point. The five terminal nodes (leaves) of the tree correspond to the five rectangular regions.

3.5.3 Two Choices at Each Step

At each step of the recursive partitioning, we face two decisions:

1. **Continue or stop?** Should we keep splitting the current region, or is it small enough (or pure enough) to become a leaf?
2. **Where to partition?** If we continue, which variable X_j should we split on, and at what cut point s?

We use a **greedy algorithm** to answer both questions. "Greedy" means that at each step we make the locally best choice without looking ahead at future splits.

For each leaf R_m, we define the following quantities:

- **Size**: $N_m = |\{x_i \in R_m\}|$ (the number of observations in the leaf).
- **Fitted value** (mean as fit): $\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$.
- **Mean squared error in leaf**: $Q_m(\mathcal{T}) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$.

3.5.4 Where to Partition

Conditional on continuing to grow, how do we determine the best split? For the j-th predictor and a cut position s, we define two half-planes:

$$R_1(j, s) = \{X \mid X_j \leq s\}, \qquad R_2(j, s) = \{X \mid X_j > s\}.$$

We search over all predictors j and all cut points s to minimize the total SSE across the two resulting leaves:

$$\min_{j,\, s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right], \tag{3.9}$$

where c_1 and c_2 are two constants. For any given split (j, s), the optimal c_1 and c_2 are simply the sample averages of y_i in the respective regions, so the minimization over (j, s) is computationally straightforward.

3.5.5 When to Stop Growing

The second decision is whether to continue splitting or to stop. A tree that keeps splitting until every leaf contains a single observation will have zero training error, but it will be massively overfit. A tree that stops too early will miss important structure. Too large $\rightarrow$ *overfitting*; too small $\rightarrow$ losing information.

The standard approach is to first grow a large tree and then *prune it back*:

1. **Step 1: Grow a big tree $\mathcal{T}_0$.** Keep splitting until some minimum node size is reached (say, each leaf has only 10 observations).
2. **Step 2: Prune**. Choose the subtree $\mathcal{T} \subset \mathcal{T}_0$ with the lowest cost function $C_\alpha(\mathcal{T})$. Here $\mathcal{T} \subset \mathcal{T}_0$ means any tree that can be obtained by collapsing any number of internal nodes in $\mathcal{T}_0$.

Cost-Complexity Pruning

$$C_\alpha(\mathcal{T}) = \sum_{m=1}^{|\mathcal{T}|} N_m \, Q_m(\mathcal{T}) + \alpha \, |\mathcal{T}|, \tag{3.10}$$

where $|\mathcal{T}|$ is the number of terminal nodes (leaves) and $\alpha \geq 0$ is a tuning parameter.

This criterion has a familiar structure: it is the total SSE (a measure of bias) plus a penalty for tree size (a measure of complexity). The parameter α determines how hard we penalize tree size. Larger α produces smaller, simpler trees. As with penalized regression, the optimal α can be chosen by cross-validation.

3.6 Random Forests

3.6.1 The Problem with a Single Tree

A single decision tree is intuitive and easy to interpret, but it has a significant weakness: high variance. Small changes in the training data can produce very different tree structures. One sample might split first on income, another on education, leading to completely different partitions. This instability is precisely the variance problem from our bias-variance decomposition.

The solution is to *average over many trees*. If individual trees are noisy but unbiased, their average will have the same bias but lower variance. This is the logic behind Random Forests (Hastie et al., 2009, Chapter 15).

3.6.2 Bagging and De-Correlation

The procedure involves two variance-reduction techniques:

Bagging (Bootstrap Aggregation). We draw B bootstrap samples from the train-

ing data (each of size N, drawn with replacement) and grow a tree T_b on each sample. The final prediction is the average:

$$\hat{f}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

Alternatively, we can use sub-sampling (draw samples of size $n' < N$ without replacement).

De-correlation. Bagging alone is not enough, because trees grown on overlapping bootstrap samples tend to be correlated. If the trees are correlated, averaging them does not reduce variance as much. To break this correlation, Random Forests add a twist: at each split, instead of searching over all p variables, we randomly select m variables (where $m \ll p$) and choose the best split among those m only. This forces different trees to use different variables, reducing their correlation.

The variance of the forest prediction is approximately:

$$V(\hat{f}) \approx \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2, \tag{3.11}$$

where ρ is the average pairwise correlation between trees and σ^2 is the variance of a single tree. The first term depends on ρ and cannot be reduced by adding more trees. The second term shrinks as B increases. To reduce $V(\hat{f})$, we can either increase B (more trees) or decrease ρ (smaller m, more de-correlation).

Random Forests

Random Forests = Tree Method + Sampling Average (Many De-correlated Trees).

The algorithm:

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a tree T_b on $\mathbf{Z}^*$: at each split, randomly select m variables from the p available, pick the best variable/split-point among those m, and split. Continue until the minimum node size $n_{\min}$ is reached.
2. Output the ensemble $\{T_b\}_{b=1}^{B}$.

Prediction: $\hat{f}_{\mathrm{RF}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

For classification: let $\hat{C}_b(x)$ be the class predicted by tree b. The forest prediction is $\hat{C}^B_{\text{RF}}(x) = \text{majority vote}\{\hat{C}_b(x)\}_{b=1}^B$.

3.6.3 Connection to Kernel Methods

Random Forests are conceptually similar to the kernel and nearest-neighbor methods we studied in Chapter 2. All of these methods make predictions by taking weighted averages of "nearby" observations. The difference lies in how "nearby" is defined:

- **Nearest neighbor and kernel methods**: the notion of proximity is fixed by a distance metric (e.g., Euclidean distance) or a bandwidth h. It does not adapt to the data.
- **Random Forests**: the notion of proximity is *adaptive*. The tree structure learns which directions in the covariate space are most informative for prediction. Two observations are "nearby" if they frequently end up in the same leaf, not if their Euclidean distance is small.

An important application of Random Forests in economics is **Causal Forests**, which we discuss next.

3.7 Causal Forests

3.7.1 Motivation: Heterogeneous Treatment Effects

In Chapter 1, we focused on estimating summary measures of the treatment effect: the Average Treatment Effect (ATE). This tells us the average impact of a policy across the entire population. But in many applications, the most interesting question is: *for whom does the treatment work best?*

A drug may cure some patients but have no effect on others. A job training program may benefit workers with low baseline skills but do little for highly skilled workers. These are *heterogeneous treatment effects* (HTEs), and they are crucial for policy design: if we know who benefits most, we can target the treatment more efficiently.

The traditional approach to detecting HTEs is subgroup analysis: split the sample by gender, age, education, etc., and estimate the treatment effect within each subgroup. The problem is that this invites *cherry picking*: running many subgroup analyses until

a statistically significant difference appears by chance. This multiple testing problem is severe when the number of potential subgroups is large.

Wager and Athey (2018) propose **Causal Forests** as a principled, data-driven solution. Causal Forests are an extension of Random Forests designed specifically to estimate heterogeneous treatment effects without the researcher pre-specifying which subgroups to examine.

3.7.2 Setup and Assumptions

We observe data (X_i, Y_i, W_i) for $i = 1, \ldots, n$, where X_i is a vector of covariates, Y_i is the outcome, and $W_i \in \{0, 1\}$ is the treatment indicator. The *Conditional Average Treatment Effect* (CATE) at a covariate value x is:

$$\tau(x) = E\left[Y_i^{(1)} - Y_i^{(0)} \mid X_i = x\right]. \tag{3.12}$$

This is the treatment effect for the subpopulation with characteristics x. It varies with x; that is what "heterogeneous" means.

Causal Forests Require Unconfoundedness

Causal Forests assume **unconfoundedness** (selection on observables):

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp\!\!\!\perp W_i \mid X_i.$$

Treatment assignment is as good as random once we condition on X_i. Therefore, *causal forests are not a method to deal with endogeneity.* If there is unobserved confounding (omitted variables that affect both treatment and outcome), causal forests will not solve the problem. They are a tool for uncovering treatment effect heterogeneity under the maintained assumption that selection on observables holds.

3.7.3 Estimation Within Leaves

The estimation logic is natural. Each tree partitions the covariate space into leaves. Within a leaf $L(x)$ containing the point x, the treatment effect is estimated by com-

paring the average outcome of treated units to the average outcome of control units:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\substack{i: W_i=1 \\ X_i \in L}} Y_i - \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\substack{i: W_i=0 \\ X_i \in L}} Y_i. \tag{3.13}$$

This is just the difference in means within the leaf, exactly the estimator we would use in a randomized experiment restricted to the subpopulation in the leaf.

The forest aggregates over many such trees by averaging the leaf-level estimates. The splitting criterion is adapted: instead of minimizing prediction error for Y, the algorithm implements the Random Forests using the criterion of *maximizing the variance of the treatment effect $\hat{\tau}(X_i)$ across groups (leaves)*. This encourages the tree to find splits where the treatment effect genuinely differs, rather than splits that merely predict the level of Y.

3.7.4 Honesty and Asymptotic Properties

A subtle but important issue arises: if the same data are used both to decide where to split and to estimate τ within each leaf, the estimates can be biased. The splits are chosen to maximize differences in treatment effects, so the within-leaf estimates will tend to overstate heterogeneity, a form of overfitting.

Wager and Athey (2018) solve this problem with **honest estimation**. A tree is *honest* if, for each training observation i, that observation is used either to decide the splits *or* to estimate τ, but not both.

The practical implementation uses **double-sample trees**:

1. Randomly divide the training sample into two halves, $\mathcal{I}$ and $\mathcal{J}$.
2. Use $\mathcal{I}$ to grow the tree (decide the split structure).
3. Use $\mathcal{J}$ to estimate τ in each leaf.

This separation ensures that the estimation step is not contaminated by the adaptive splitting decisions. Wager and Athey (2018) show that honest causal forests are *consistent* and *asymptotically normal*, which means we can construct valid confidence intervals for the CATE $\tau(x)$, a remarkable result for a machine learning method.

3.7.5 Application: Social Media and Polarization

Levy (2021) studies the effect of social media on news consumption and political polarization using a field experiment. The main experiment randomly assigns Facebook users to be shown counter-attitudinal news content. Causal forests are used (see Online Appendix C.5 of the paper) to explore heterogeneous treatment effects, that is, to identify which subgroups of participants are most affected by exposure to content from the other side. This is a natural application: the experimenter cannot pre-specify all relevant subgroups, and causal forests let the data reveal which characteristics matter for treatment effect heterogeneity.

3.8 Neural Networks

3.8.1 Overview

The third family of machine learning methods we introduce is **neural networks**. Neural networks have attracted enormous public attention in recent years, powering applications ranging from ChatGPT to AlphaGo. The underlying mathematics, however, is remarkably straightforward.

3.8.2 Architecture of a Single-Hidden-Layer Network

Consider a prediction problem where we observe inputs $X = (X_1, \ldots, X_p)$ and want to predict an output Y. A single-hidden-layer feed-forward neural network introduces an intermediate set of *hidden units* $Z = (Z_1, \ldots, Z_M)$ that transform the inputs before they are used to predict Y. The model has three layers:

- **Input layer**: the covariates X.
- **Hidden layer**: the derived features $Z_1, \ldots, Z_M$.
- **Output layer**: the predicted values $Y_1, \ldots, Y_K$.

Step 1: From inputs to hidden units. Each hidden unit $m = 1, \ldots, M$ computes a linear combination of the inputs and then passes it through a non-linear *activation function* σ:

$$Z_m = \sigma\big(\alpha_{0m} + \alpha_m^T X\big), \qquad m = 1, \ldots, M. \tag{3.14}$$

The activation function $\sigma(\cdot)$ is typically the logistic function $\sigma(v) = 1/(1+e^{-v})$, a step

function, or the popular ReLU function $\sigma(v) = \max(0, v)$. This non-linearity is critical: it is what distinguishes a neural network from a linear model. If σ is a linear function, the composition of two linear transformations would be another linear transformation, and the hidden layer would be redundant.

Step 2: From hidden units to outputs. Each output unit $k = 1, \ldots, K$ computes:

$$T_k = \beta_{0k} + \beta_k^T Z, \qquad Y_k = f_k(X) = g_k(T), \quad k = 1, \ldots, K, \tag{3.15}$$

where $g(\cdot)$ is another non-linear function (e.g., Logit model).

Figure 3.4 illustrates the architecture of a single-hidden-layer feed-forward neural network.

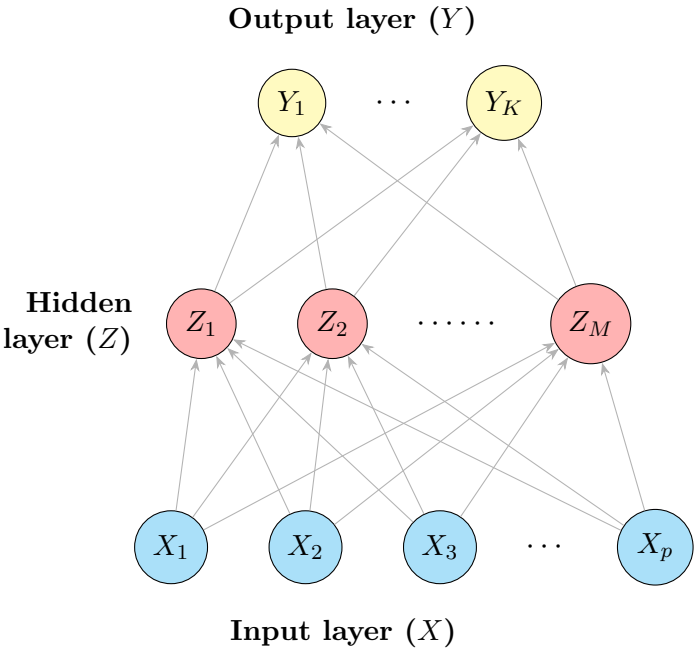


Figure 3.4: Schematic of a single-hidden-layer, feed-forward neural network. Input features X are transformed into hidden units Z via non-linear activation functions, then combined linearly to produce outputs Y.

3.8.3 Why "Neural" Networks?

The name reflects the original motivation: modeling the human brain. Each unit represents a *neuron*, and each weighted connection represents a *synapse*. When a step function is used as the activation function, a neuron "fires" (outputs 1) when the total

input signal $\alpha_{0m} + \alpha_m^T X$ exceeds a threshold, mimicking the way biological neurons fire when stimulated beyond a certain level.

There can be multiple hidden layers, giving rise to *deep* neural networks. The depth of the network is the number of hidden layers, and it is this depth that gives "deep learning" its name. Each additional layer allows the network to learn increasingly abstract representations of the data.

3.8.4 Estimation and Regularization

The parameters $\theta = (\alpha, \beta)$ are estimated by minimizing a loss function, namely nonlinear least squares for regression.

With many hidden units and many layers, neural networks have an enormous number of parameters. A single-hidden-layer network with 100 inputs and 50 hidden units already has $100 \times 50 + 50 = 5{,}050$ parameters in the first layer alone. To prevent overfitting, we **regularize** by adding a penalty term, exactly as in Ridge regression:

Neural Network Regularization

$$\min_{\theta} \; R(\theta) + \lambda J(\theta), \qquad J(\theta) = \sum_{k,m} \beta_{km}^2 + \sum_{m,l} \alpha_{ml}^2, \tag{3.16}$$

where $R(\theta)$ is the loss function (e.g., SSR) and λ is the tuning parameter controlling the strength of regularization.

The penalty $J(\theta)$ is the sum of squared weights throughout the network, the neural network analogue of the Ridge penalty. It prevents any single connection from becoming too large, which would allow the network to memorize noise. As always, λ is chosen by cross-validation.

3.9 Conclusion and Caveats

Let us step back and take stock of what we have learned.

The bias-variance tradeoff is the organizing principle of this chapter. Model complexity is double-edged: too simple and we miss the signal (high bias); too complex and we memorize the noise (high variance). Every model selection method is, at its core, a strategy for navigating this tradeoff.

Goodness-of-fit criteria (adjusted R^2, AIC, BIC, and cross-validation) provide the standards by which we evaluate models. Among these, cross-validation is particularly valuable because it directly measures out-of-sample prediction quality.

Machine learning algorithms give us automatic procedures for searching the model space. Penalized regression (LASSO, Ridge, Elastic Net) stays within the linear model and controls complexity by shrinking coefficients. Tree-based methods (CART, Random Forests) partition the covariate space and average within regions. Neural networks compose non-linear transformations to learn flexible representations.

Causal Forests are an important application for economists: they adapt Random Forests to estimate heterogeneous treatment effects under unconfoundedness, providing a principled alternative to ad hoc subgroup analysis.

Machine Learning Is Not a Substitute for Economic Reasoning

Remember: these are *statistical tools* for prediction and model selection. The most important ingredient in empirical economics is still your **economic intuition**. Machine learning can tell you which variables predict Y well, but it cannot tell you which variables have a *causal* effect on Y. Never let a data-driven algorithm override substantive economic reasoning. For example, you should never exclude *education* from a wage equation just because AIC or BIC tells you to do so; economic theory demands its inclusion.

One final observation. In this chapter, we focused on model selection conditional on the *unconfoundedness* assumption. Everything we discussed (penalized regression, trees, forests, neural networks) is about prediction: how to best approximate $E[Y \mid X]$. We did not use any prior knowledge about causal structure.

But in economics, we often *do* have such knowledge. We know that education causally affects wages. We know that prices causally affect demand. Can we incorporate this knowledge into the model selection process? The answer is yes, and the tool for doing so is the **causal graph** (Directed Acyclic Graph, or DAG). This will be the topic of Chapter 4.

Homework Problems

1. **Bias-variance decomposition.** Consider the data-generating process $Y = f(X) + \varepsilon$ with $E[\varepsilon] = 0$ and $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$. Let $\hat{f}(x)$ be an estimator trained on a random sample. Derive the bias-variance decomposition:

$$E\Big[(Y - \hat{f}(x_0))^2\Big] = \sigma_\varepsilon^2 + \Big[\text{Bias}(\hat{f}(x_0))\Big]^2 + \text{Var}(\hat{f}(x_0)).$$

   Clearly state where you use the assumption that ε is independent of $\hat{f}$.

2. **LASSO vs. Ridge.** Consider the penalized regression problem:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \left(\|\beta\|_p\right)^p.$$

   (a) Explain intuitively why the ℓ_1 penalty ($p = 1$) can produce exact zeros in $\hat{\beta}$ while the ℓ_2 penalty ($p = 2$) cannot.

   (b) In a wage regression with 50 potential covariates, would you prefer LASSO or Ridge? Justify your choice in terms of interpretability and the economic context.

3. **Cross-validation for polynomial order selection.** Suppose you have $n = 100$ observations from the DGP $Y = 1 + 2X - 0.5X^2 + \varepsilon$, $\varepsilon \sim N(0, 4)$. Describe how you would use 5-fold cross-validation to choose between a linear, quadratic, and cubic polynomial model. Which model do you expect to be selected, and why?

Practice Example: LASSO and Ridge Regression in Stata

This practice example illustrates penalized regression methods in Stata. We simulate a dataset with a known sparse data-generating process (only a few covariates truly matter) and compare the performance of OLS, LASSO, and Ridge regression.

Setup

The data-generating process is:

$$Y_i = 2X_{1i} + 3X_{2i} - 1.5X_{3i} + \varepsilon_i, \qquad \varepsilon_i \sim N(0,1),$$
$$X_{ji} \sim N(0,1) \text{ for } j = 1, \ldots, 20, \qquad n = 200.$$

Only three of the twenty covariates have nonzero coefficients. The remaining seventeen are pure noise variables with $\beta_j = 0$. This "sparse" structure is the kind of problem where LASSO is expected to beat OLS out of sample, because LASSO can set the irrelevant coefficients to exactly zero. In sample, OLS always fits best (it minimizes the in-sample sum of squared residuals by construction), so the do-file also simulates a fresh holdout sample of 200 observations from the same DGP, never used in estimation, and evaluates every fitted model on it.

Learning Objectives

1. OLS uses all 20 regressors and assigns nonzero (noisy) coefficients to the irrelevant variables.
2. LASSO identifies the three relevant variables by shrinking the irrelevant coefficients to zero.
3. Ridge shrinks all coefficients toward zero but does not set any exactly to zero.
4. Cross-validation selects the tuning parameter λ that minimizes out-of-sample prediction error.
5. In-sample and out-of-sample RMSE comparison: OLS, LASSO, Ridge, and the irreducible noise $\sigma = 1$. In the simulated data OLS has the lowest in-sample RMSE (0.93, below the realized noise RMSE of 0.99, so it fits part of the noise), but on the holdout sample LASSO has the lowest RMSE (1.06, against 1.12 for OLS, 1.11 for Ridge, and 1.04 for the irreducible noise). Ridge shrinks only lightly here, so it behaves much like OLS.

Replication Code

The Stata do-file `src/chapter 3/ch3_ml.do` reproduces all results. The simulated dataset is saved to `data/chapter 3/ch3_ml.dta`.

Chapter 4

Directed Acyclic Graphs

4.1 Introduction

Causal inference is the central topic of applied economics. Throughout this book, we have relied almost exclusively on the *potential outcome* (PO) framework, proposed by Rubin (1974) and developed extensively by Imbens and Rubin (2015) and Angrist and Pischke (2009). This framework is sometimes called the "Rubin Causal Model," and it has become the dominant language of causal inference in economics.

But is this the only statistical framework for dealing with causal questions? Of course not. *Graphical models*, and in particular *Directed Acyclic Graphs* (DAGs), constitute another important and influential approach to causality. This framework is most closely associated with the work of Judea Pearl, an Israeli-American computer scientist and philosopher who was awarded the 2011 Turing Award "for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning." Pearl's approach has deep roots in computer science and artificial intelligence, and it provides a strikingly different, yet complementary, perspective on how to think about cause and effect.

In this chapter, we introduce the DAG framework and its key concepts: graphs, Bayesian networks, causal edges, d-separation, the *do*-operator, and the backdoor and frontdoor adjustment formulas. We then turn to a critical comparison of DAGs and potential outcomes, drawing heavily on Imbens (2020), who provides a thoughtful and balanced assessment of where each framework excels and where it falls short. The

goal is not to declare a winner, but to understand what DAGs can offer to applied economists, and what limitations remain.

Before diving in, it is worth noting the key references for this chapter:

- Pearl (2009), *Causality*, Cambridge University Press: the definitive technical treatment.
- Pearl and Mackenzie (2018), *The Book of Why*, Allen Lane: a highly accessible introduction for a general audience.
- Neal (2020), *Introduction to Causal Inference*: an excellent online course and textbook, freely available online.[1]
- Imbens (2020): a critical and constructive assessment from an economist's perspective.

4.2 The DAG Approach: Graphs

We begin with the mathematical object at the heart of this framework: the *graph*.

Definition 4.2.1 (Graph)**.** A **graph** is a collection of **nodes** (vertices) and **edges** that connect them. Two nodes are called **adjacent** if they are connected by an edge.

There are two basic types of graphs. In an *undirected graph*, edges have no direction. They simply indicate that two nodes are connected. In a *directed graph*, each edge has a direction: it goes out of one node (the **parent**) and into another (the **child**).

[1] https://www.bradyneal.com/causal-inference-course

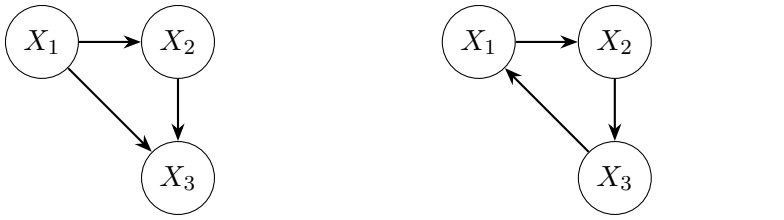
(a) Undirected Graph

(b) Directed Graph

Figure 4.1: Undirected and directed graphs.

A **path** is any sequence of adjacent nodes, regardless of the direction of the edges. A **directed path** is a path where all edges point in the same direction, from the first node toward the last. If there is a directed path from X to Y, we say X is an **ancestor** of Y, and Y is a **descendant** of X.

Definition 4.2.2 (Directed Acyclic Graph)**.** A **directed acyclic graph (DAG)** is a directed graph that contains no cycles. That is, there is no directed path that starts and ends at the same node.



(a) No Cycle (DAG)

(b) Cycle (Not a DAG)

Figure 4.2: A directed acyclic graph and a directed graph with a cycle.

The "acyclic" restriction is crucial. It rules out feedback loops: if X causes Y and Y

causes X, there is a cycle, and the graph is not a DAG. As we will discuss later, this is one of the key limitations of the DAG framework for economics: many economic models involve equilibrium and simultaneity, which are inherently cyclical.

4.3 Bayesian Networks

How do we connect graphs to statistics? The first step is to connect graphs to probability distributions. This is the idea behind *Bayesian networks*.

4.3.1 From Joint Distributions to Factorization

Recall that any joint distribution can be decomposed using the chain rule of probability:

$$P(x_1, x_2, \ldots, x_n) = P(x_1) \prod_{i \neq 1} P(x_i \mid x_{i-1}, \ldots, x_1). \tag{4.1}$$

For example, with three variables:

$$P(x_1, x_2, x_3) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2, x_1).$$

This decomposition is always valid. It follows from the definition of conditional probability and involves no assumptions. But it is also unwieldy: every variable depends on *all* preceding variables.

The key insight of Bayesian networks is that we can *simplify* this factorization if we are willing to make assumptions about the dependence structure. For instance, if $X_1 \perp X_3 \mid X_2$, then $P(x_3 \mid x_2, x_1) = P(x_3 \mid x_2)$, and the factorization simplifies to:

$$P(x_1, x_2, x_3) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2).$$

By making more independence assumptions, we can simplify further. The idea is to use a graph to *represent* the assumed dependence structure, creating a one-to-one mapping between the graph G and the probabilistic relations in P.

4.3.2 The Minimality Assumption and Bayesian Network Factorization

Assumption 4.3.1 (Minimality)**.**

1. **Local Markov Assumption:** Given its parents in the DAG, a node X is independent of all its non-descendants.
2. **Minimal Independence:** Adjacent nodes in the DAG are dependent.

The Local Markov Assumption says that the dependence structure is "local": once you know a node's parents, you do not need to know anything else about its non-descendants.

Theorem 4.3.1 (Bayesian Network Factorization)**.** *Given a probability distribution P and a DAG G satisfying the Local Markov Assumption, P factorizes according to G as:*

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i \mid \text{pa}_i) \tag{4.2}$$

where pa_i *denotes the set of parents of node i in the graph G.*

The Bayesian Network Factorization tells us something powerful: if the probability distribution P has the statistical structure shown in graph G, then each variable x_i depends *only* on its parents pa_i in the graph. We say that "G represents P," or equivalently that "G and P are compatible," or that "P is Markov relative to G."

4.3.3 A Worked Example

Suppose we have four variables x_1, x_2, x_3, x_4. The full chain-rule decomposition is:

$$P(x_1, x_2, x_3, x_4) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2, x_1) \cdot P(x_4 \mid x_3, x_2, x_1).$$

Now suppose our DAG is:

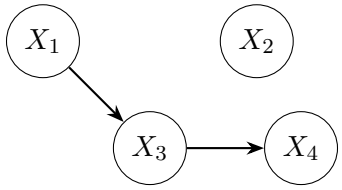


Figure 4.3: A DAG with four variables, where X_3 depends on X_1 and X_4 depends on X_3.

In this graph, X_3 has one parent (X_1), X_4 has one parent (X_3), and X_1, X_2 have no parents. By the Bayesian Network Factorization, we can simplify the full decomposition as:

$$P(x_1, x_2, x_3, x_4) = P(x_1) \cdot P(x_2) \cdot P(x_3 \mid x_1) \cdot P(x_4 \mid x_3). \tag{4.3}$$

The edges in the graph encode exactly which statistical dependencies exist.

4.4 Basics of Causal Graphs

4.4.1 Causal Graphs

Up until now, we have discussed only *statistical* dependencies. The arrows in the graphs above encode conditional independence relations, but they do not necessarily carry a causal interpretation. To move from statistical associations to causal claims, we need one more assumption.

Assumption 4.4.1 (Causal Edges)**.** In a directed graph, every parent is a **direct cause** of all its children.

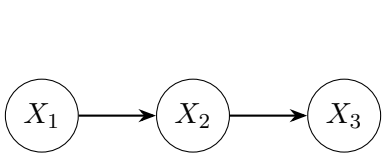
By adding this assumption, the arrows in the DAG acquire a causal meaning. Directed paths correspond to *causation*: if there is a directed path from T to Y, then T (possibly indirectly) causes Y. The DAG now represents not just statistical dependencies, but a full *causal model*.

A more mathematically rigorous formulation of this idea uses the framework of *structural equation models* (SEMs), where each variable is written as a function of its parents and an independent error term.

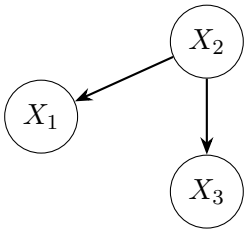
4.4.2 Graphical Building Blocks

There are three fundamental building blocks in any DAG. Understanding these three structures is the key to understanding how association and causation flow through a causal graph.

4.4.2.1 Chains, Forks, and Colliders



(a) Chain



(b) Fork

Figure 4.4: Chain and fork structures.

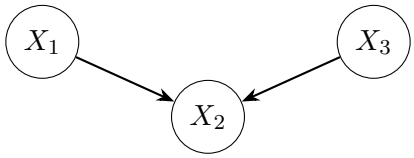


Figure 4.5: Collider (immorality) structure.

Two key principles govern the flow of information through these structures:

- **Association is symmetric:** X_1 and X_3 are associated in both chains and forks (but *not* in colliders).
- **Causation is asymmetric:** in the chain, X_1 causes X_3 (through X_2), but X_3 does not cause X_1.

4.4.2.2 Conditioning Blocks Chains and Forks

In chains and forks, conditioning on the middle variable X_2 *blocks* the flow of association between X_1 and X_3.

Proposition 4.4.1. *In the chain* $X_1 \to X_2 \to X_3$*:*

$$P(x_1, x_3 \mid x_2) = P(x_1 \mid x_2) \cdot P(x_3 \mid x_2).$$

That is, $X_1 \perp X_3 \mid X_2$*.*

Proof. By Bayesian factorization and the Local Markov Assumption:

$$P(x_1, x_2, x_3) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2).$$

Therefore:

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

□

The same result holds for forks: if X_2 is a common cause of X_1 and X_3, then conditioning on X_2 renders X_1 and X_3 independent. Intuitively, the "spurious" association between X_1 and X_3 is entirely due to their shared cause X_2; once we know X_2, there is no residual association.

4.4.2.3 Conditioning Opens Colliders

Colliders behave in exactly the opposite way. In the collider $X_1 \to X_2 \leftarrow X_3$, the variables X_1 and X_3 are *marginally independent*:

$$\begin{aligned} P(x_1, x_3) &= \int_{x_2} P(x_1) \cdot P(x_3) \cdot P(x_2 \mid x_1, x_3)\, dx_2 \\ &= P(x_1) \cdot P(x_3) \int_{x_2} P(x_2 \mid x_1, x_3)\, dx_2 = P(x_1) \cdot P(x_3). \end{aligned} \tag{4.4}$$

But *conditioning on* the collider X_2 *creates* a dependence between X_1 and X_3:

$$P(x_1, x_3 \mid x_2) \neq P(x_1 \mid x_2) \cdot P(x_3 \mid x_2) \quad \text{in general.}$$

Collider Bias: The Bad Control Problem

Consider a simple example. Let X_1 be "good-looking," X_3 be "kindness," and X_2 be "marital status." Suppose being good-looking and being kind both independently make someone more likely to be married ($X_1 \to X_2 \leftarrow X_3$).
Among unmarried people (conditioning on $X_2 = 0$), you observe a *negative* correlation between attractiveness and kindness: "Why are all the good-looking ones so unkind?" This is pure selection bias. It arises because people who are both good-looking and kind already get married. You are conditioning on a collider,

or a post-determined outcome (marital status). There is no causal relationship between attractiveness and kindness.
This is precisely the **bad control problem** in econometrics: controlling for a variable that is itself affected by the treatment (or by the treatment's causes) can introduce bias rather than remove it.

Remark 4.4.1. The homework at the end of this chapter asks you to prove formally that conditioning on the collider X_2 creates a dependence between X_1 and X_3.

4.4.3 Blocked Paths and d-Separation

The three building blocks (chains, forks, and colliders) give us the vocabulary to describe when and how association flows through a DAG. The general concept is *d-separation*.

Definition 4.4.1 (Blocked Path)**.** A path between X and Y is **blocked** by a conditioning set Z if **either** of the following is true:

1. Along the path, there is a chain ($\rightarrow W \rightarrow$) or a fork ($\leftarrow W \rightarrow$) where $W \in Z$.
2. There is a collider W on the path such that neither W nor any descendant of W is in Z.

The first condition says: conditioning on a non-collider blocks the path. The second condition says: *not* conditioning on a collider (or its descendants) also blocks the path. Association flows along unblocked paths and does *not* flow along blocked paths.

Definition 4.4.2 (d-Separation)**.** Two sets of nodes X and Y are **d-separated** by a set of nodes Z if *all* paths between any node in X and any node in Y are blocked by Z.

d-Separation is the graphical criterion for conditional independence. When X and Y are d-separated by Z, all association flows between X and Y are blocked by Z.

Theorem 4.4.1 (d-Separation and Statistical Independence, Theorems 1.2.4–1.2.5 in Pearl, 2009; Theorem 3.1 in Neal, 2020)**.** *If X and Y are d-separated in a DAG G conditional on Z, then X and Y are independent conditioned on Z in every distribution compatible with G:*

$$X \perp_G Y \mid Z \;\Rightarrow\; X \perp_P Y \mid Z, \quad \forall\, P \text{ compatible with } G.$$

Conversely, if X and Y are independent conditional on Z in all P compatible with G, then X and Y are d-separated in G conditional on Z:

$$(\forall\, P \text{ compatible with } G,\ X \perp_P Y \mid Z) \ \Rightarrow\ X \perp_G Y \mid Z.$$

This theorem is a *bridge* between graphs and probability. It tells us exactly how to translate statistical independence statements into graphical statements, and vice versa.

The notion of d-separation leads directly to the question of causal identification. Recall:

- **Associations** flow along all unblocked paths.
- **Causation** flows along *directed* unblocked paths.

To isolate the causal effect of T on Y, we need to ensure that there is no *non-causal* association between T and Y. Graphically, this means: in the manipulated graph $G_{\underline{T}}$ (the graph with all outgoing edges from T removed), T and Y should be d-separated by the control set W. In other words, all non-causal ("backdoor") paths must be blocked.

4.4.4 The *do*-Operator

The DAG framework formalizes causal interventions using the *do*-operator, introduced by Pearl (2009). We define $\text{do}(T = t)$ as an intervention that forces the entire population to receive treatment level t.

In terms of potential outcomes, the *do*-operator corresponds to:

$$P(y \mid \text{do}(t)) = P(Y = y \mid \text{do}(T = t)) = P(Y(t) = y), \tag{4.5}$$

where $Y(t)$ is the potential outcome under treatment t, exactly as defined in Chapter 1.

The notation $P(y \mid \text{do}(t))$ is *not* the same as the conditional probability $P(y \mid t)$. The latter conditions on *observing* $T = t$, which may reflect both causation and selection. The former represents what would happen if we *intervened* to set $T = t$, breaking all arrows pointing into T.

Definition 4.4.3 (Identifiability via the *do*-Operator)**.** A causal quantity Q is **identifiable** if we can reduce an expression involving $\text{do}(\cdot)$ to one that involves only standard (do-free) probabilities of observed variables. This is precisely analogous to the PO framework, where identification means reducing expressions involving potential out-

comes $Y(t)$ to expressions involving only observed data.

4.5 Causal Identification Method in Causal Graphs

4.5.1 Backdoor Adjustment

One of the most important identification results in the DAG framework is the *backdoor adjustment*. Non-directed unblocked paths from T to Y, paths that go "through the back" of T, are called **backdoor paths**. These paths create non-causal associations between treatment and outcome.

Definition 4.5.1 (Backdoor Criterion)**.** A set of variables W satisfies the **backdoor criterion** relative to (T, Y) if:

1. W blocks all backdoor paths from T to Y; and
2. W does not contain any descendant of T.

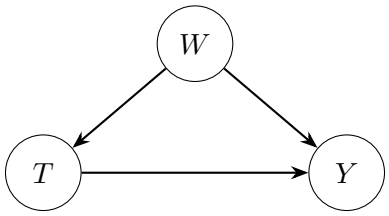


Figure 4.6: A backdoor path $T \leftarrow W \rightarrow Y$ through the confounder W.

Theorem 4.5.1 (Backdoor Adjustment)**.** *If W satisfies the backdoor criterion relative to (T, Y), then:*

$$P(y \mid do(t)) = \int_w P(y \mid t, w)\, P(w)\, dw. \tag{4.6}$$

This should look familiar. The variable set W is exactly what we usually call "control variables" in applied economics. The backdoor criterion is the graphical analogue of the "selection on observables" (or "unconfoundedness") assumption from the potential outcome framework (Chapter 1, Section 1.5). The backdoor adjustment formula is just the familiar formula for the average treatment effect under unconfoundedness:

$$\text{ATE} = E[Y(1)] - E[Y(0)] = \int \{E[Y \mid T = 1, W = w] - E[Y \mid T = 0, W = w]\}\, dF_W(w).$$

4.5.2 Frontdoor Adjustment

One of the most interesting identification strategies unique to the DAG framework is the *frontdoor adjustment*. This idea has no direct analogue in the standard PO toolkit, and it illustrates the creative power of graphical reasoning.

4.5.2.1 Setup

Consider the following DAG, where W is an *unobserved* confounder:

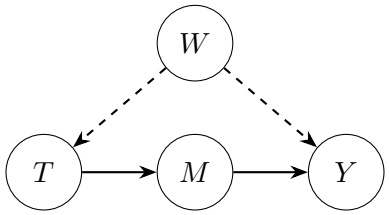


Figure 4.7: Frontdoor structure: T affects Y only through the mediator M, with W an unobserved confounder (dashed).

In this graph, T affects Y *only* through the mediator M. The confounder W creates a backdoor path $T \leftarrow W \rightarrow Y$, so we cannot simply regress Y on T. But we can identify the causal effect in three steps:

1. **Identify the effect of T on M:** There is no backdoor path from T to M (the only path is the direct arrow $T \rightarrow M$), so $P(m \mid \text{do}(t)) = P(m \mid t)$.
2. **Identify the effect of M on Y:** There is a backdoor path from M to Y through W, but it can be blocked by conditioning on T. So $P(y \mid \text{do}(m)) = \sum_{t'} P(y \mid m, t')P(t')$.
3. **Combine:** Chain the two results together.

Definition 4.5.2 (Frontdoor Criterion)**.** A set of variables M satisfies the **frontdoor criterion** relative to (T, Y) if:

1. M completely mediates the causal effect of T on Y (i.e., all directed paths from T to Y go through M).
2. There is no unblocked backdoor path from T to M.
3. All backdoor paths from M to Y are blocked by T.

Theorem 4.5.2 (Frontdoor Adjustment)**.** *If T, M, Y satisfy the frontdoor criterion,*

then:

$$P(y \mid do(t)) = \sum_{m} P(m \mid t) \sum_{t'} P(y \mid m, t')\, P(t'). \tag{4.7}$$

This is a remarkable result: we can identify the causal effect of T on Y *even though there is an unobserved confounder*, as long as we observe a complete mediator M.

4.5.2.2 Limitations

The frontdoor adjustment is intellectually beautiful, but its practical applicability is limited. The requirement that M *completely* mediates the effect of T on Y is very strong. In most economic settings:

- What if T affects Y through other channels besides M?
- What if some unobserved U affects both M and Y?
- What if the confounder W also affects M directly?

Any of these violations breaks the frontdoor criterion. Finding a genuine complete mediator in real economic applications is extremely difficult. How to apply this method to economics remains an open question.

4.5.3 Non-parametric Identification: *do*-Calculus

The backdoor and frontdoor criteria are *sufficient* conditions for causal identification, but they are not *necessary*. That is, there may be causal effects that are identifiable even though neither criterion is satisfied. Can we find a *complete* set of rules that determines whether any causal effect is identifiable in any DAG?

The answer is yes: the *do*-calculus, developed by Pearl (2009).

4.5.3.1 Notation

We need some notation for manipulated graphs:

- $G_{\overline{X}}$: take graph G and *remove all incoming edges* to X.
- $G_{\underline{X}}$: take graph G and *remove all outgoing edges* from X.
- $Z(W)$: the set of nodes of Z that are not ancestors of any node of W in $G_{\overline{T}}$.

4.5.3.2 The Three Rules

Theorem 4.5.3 (Rules of *do*-Calculus, Pearl, 2009). *1.* ***Rule 1 (Deletion of variable):***

$$P(y \mid do(t), z, w) = P(y \mid do(t), w) \quad \text{if } Y \perp_{G_{\overline{T}}} Z \mid T, W.$$

2. ***Rule 2 (do-variable exchange):***

$$P(y \mid do(t), do(z), w) = P(y \mid do(t), z, w) \quad \text{if } Y \perp_{G_{\overline{T}\underline{Z}}} Z \mid T, W.$$

3. ***Rule 3 (Deletion of action):***

$$P(y \mid do(t), do(z), w) = P(y \mid do(t), w) \quad \text{if } Y \perp_{G_{\overline{T}\overline{Z(W)}}} Z \mid T, W.$$

4.5.3.3 Intuition Behind the Rules

What do these three rules mean intuitively?

Rule 1 is an extension of d-separation. If we erase the do-operator from do(t), the condition becomes $Y \perp_G Z \mid W$, just standard conditional independence under the Markov assumption. In words: if Z provides no additional information about Y once we know T and W (in the manipulated graph), then we can drop Z from the expression.

Rule 2 is an extension of the backdoor adjustment. If we erase do(t), the condition becomes: W blocks all non-causal paths between Z and Y in the graph where outgoing edges from Z are removed. In other words, we can replace do(z) with simply observing z, because W already controls for all confounding between Z and Y.

Rule 3 says that if, in the appropriate manipulated graph, Z and Y are d-separated, then the intervention do(z) has no effect on Y and can be dropped entirely.

Theorem 4.5.4 (Completeness of *do*-Calculus, Pearl, 2009). *A causal effect Q is identifiable in a model characterized by a graph G if and only if there exists a finite sequence of transformations, each conforming to one of Rules 1, 2, or 3, that reduces Q into a standard (do-free) probability expression involving observed quantities.*

The do-calculus is *complete*: these three rules are sufficient to identify *every* identifiable causal effect. If a causal effect cannot be identified using these rules, then it is genuinely not identifiable from observational data given the assumed graph.

Non-parametric Only

An important caveat: the do-calculus concerns *non-parametric* identification only. It asks whether the causal effect can be expressed in terms of the observed joint distribution without any functional form assumptions. In many economic applications, parametric or semi-parametric assumptions (e.g., linearity, monotonicity) provide additional identifying power that the do-calculus does not capture.

4.5.4 An Example: Returns to College

Let us consider a concrete example. Suppose we are interested in the causal effect of college education (T) on wages (Y). A plausible DAG might include:

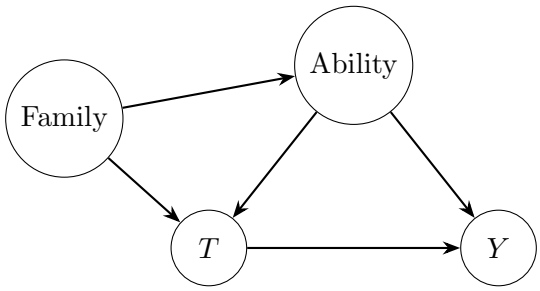


Figure 4.8: A causal DAG for college education and wages, with Ability and Family as background factors.

The question is: which variables do we need to control for? The DAG framework gives us a systematic answer through the backdoor criterion. The backdoor paths from T to Y are:

- $T \leftarrow \text{Ability} \rightarrow Y$
- $T \leftarrow \text{Family} \rightarrow \text{Ability} \rightarrow Y$

Controlling for Ability blocks both paths. Controlling for Family alone blocks only the second path but not the first (since Ability remains an unblocked non-collider). The DAG tells us precisely what we need to control for, and what we should *not* control for.

4.6 DAGs in Economics: Pros and Cons

Having introduced the DAG framework, we now turn to the central question of this chapter: what can DAGs offer to applied economists, and where do they fall short? We follow the structure of Imbens (2020), who provides a balanced and thoughtful assessment.

4.6.1 Advantages of DAGs

Pro 1: Clarity of Illustration. DAGs make causal assumptions *explicit and visual*. Consider two of the most common identification strategies in economics:

Unconfoundedness. The assumption that treatment is independent of potential outcomes conditional on observables $(Y(0), Y(1) \perp T \mid W)$ can be illustrated by a simple DAG: $W \to T$, $W \to Y$, $T \to Y$. The backdoor path $T \leftarrow W \to Y$ is blocked by conditioning on W. This is immediately transparent in the graph.

Instrumental variables. The IV assumptions, relevance (Z affects T), exclusion (Z does not directly affect Y), and independence (Z is as good as randomly assigned), correspond to a DAG: $Z \to T \to Y$, with no direct edge $Z \to Y$ and no common cause of Z and Y. The exclusion restriction, which can be hard to state precisely in words, becomes visually obvious: there is simply no arrow from Z to Y.

Pro 2: Tool for Analyzing Complicated Causal Models. When the causal model involves many variables with complex interdependencies, DAGs provide systematic machinery (d-separation, do-calculus) for determining which causal effects are identifiable and which control sets are valid. This is the setting where DAGs truly shine: with a large number of variables, it becomes very difficult to track all the possible confounding paths by hand, and the graphical approach provides an algorithmic solution.

Pro 3: Frontdoor Criterion. As discussed earlier, the frontdoor adjustment provides an identification strategy that has no direct analogue in the PO framework. While its practical applicability remains limited (complete mediators are rare in economics), it demonstrates the creative power of graphical reasoning and remains an active area of research.

Pro 4: Systematic Analysis of Mediation Effects. DAGs can shed light on the identification of mediation (or mechanism) effects in a rigorous way. The standard

"mediation effect test" (the Baron-Kenny approach) implicitly assumes a very simple causal structure, one that rules out any unobserved confounders between the treatment and the mediator, between the mediator and the outcome, or between the treatment and the outcome.

The Mediation Test Is Often a Regression Monkey Exercise

The standard mediation test requires you to accept a very simple causal structure just to implement an off-the-shelf test. But think about what this structure implies:

- What if there is an unobserved U affecting both T and Y?
- What if there is an unobserved U affecting both M and T?
- What if there is an unobserved U affecting both M and Y?

If you truly believe this simple structure (no confounders anywhere, full mediator), then you can directly identify the effect of T on Y using OLS or the frontdoor method! It means that when you use research design (DID, IV etc.) to rule out endogeneity issues but at the same time want to implement a mediation test, you are doing research in an **inconsistent** way. If you think there are confounders, then you cannot run simple mediation test.

For most economists, the standard mediation test is, "a typical behavior of regression monkey." It is a super-simplified reduced DAG with very strong assumptions. Meanwhile, it leads to inconsistency when the authors use research designs to rule out confounders. We should think about mediation in a more general DAG framework.

The standard mediation test originates from psychology, where RCTs provide a clean environment without confounders. However, this is not the case in economics where we always struggle with observational data in real life. Therefore, in economics, standard mediation test can only give us some suggestive evidence at best. DO NOT put too much importance on the results of these tests.

Generally speaking, do not use off-the-shelf mediation test without thinking by yourself!

Then how *should* we investigate mechanisms in empirical research? I suggest several approaches:

1. Use **heterogeneity analysis** to provide indirect evidence of mechanisms.
2. Think from a **general DAG framework**: construct a causal graph based on

your knowledge of the economics and test the implied relationships.

3. Construct a **theoretical model**, derive comparative statics, transform them into testable regressions.

4. Construct a **structural model**, illustrate the mechanism through decomposition and counterfactual analysis.

4.6.2 Disadvantages of DAGs

Con 1: DAGs Require Ex Ante Causal Structure. The DAG approach develops powerful machinery for identification, but it requires the whole structure of the causal model. Little is said about *why* we should believe a particular structure (the "before model" question) or about estimation and inference (the "after model" question). As Imbens (2020) notes, the DAG literature sometimes appears to offer "a set of solutions in search of problems."

For economists, both the *before* and *after* are crucial. We want to know where the model comes from (economic theory? institutional knowledge?) and how to estimate the causal effect with valid inference in finite samples.

Con 2: DAGs Do Not Fit IV Very Well. The instrumental variables setting reveals several limitations of the DAG approach:

- **Shape restrictions:** The monotonicity assumption (central to the LATE theorem of Imbens and Angrist, 1994) cannot be naturally expressed in a DAG. DAGs encode conditional independence relations, not shape restrictions on functional forms.
- **Heterogeneity:** The LATE theorem is not easily derived in a DAG framework. The PO framework, with its explicit notation for heterogeneous potential outcomes $Y_i(0), Y_i(1)$, handles treatment effect heterogeneity much more naturally.
- **RDD:** Similarly, the regression discontinuity design does not gain much from the DAG approach.

More generally, the inability to fit IV well reveals two deeper weaknesses: DAGs are not convenient for expressing economic structural assumptions, and the PO framework can deal with heterogeneity issues better.

Con 3: DAGs Cannot Capture Simultaneity and Equilibrium. By definition, a DAG is *acyclic*, so it cannot contain feedback loops. But many economic models are built on equilibrium concepts that are inherently cyclical: supply determines price, and price determines supply. DAGs cannot naturally express simultaneity.

Imbens (2020) attempts to represent a supply-and-demand equilibrium in a DAG, but the result is "not so successful": the time-indexed expansion required to avoid cycles is awkward and loses the economic content of the simultaneity.

4.6.3 Why Is PO Still More Popular in Economics?

Summarizing Imbens (2020), there are five features of the PO framework that explain its dominance in applied economics:

1. Some common economic assumptions (monotonicity, convexity) are easily captured in PO but not in DAGs.
2. Potential outcomes connect naturally to traditional economic models (supply and demand, utility maximization).
3. Applied economics typically involves models with relatively few variables, where the systematic machinery of d-separation adds less value.
4. PO handles treatment effect heterogeneity (ATE, ATT, LATE) more naturally.
5. PO connects closely to the implementation of the method and its inference: estimation, standard errors, confidence intervals are integral parts of the framework.

A final, pragmatic point: the DAG literature has not provided substantive empirical examples demonstrating clear advantages over the PO approach. Most examples from Pearl (2009) are "toy models." Until concrete applied examples emerge, adoption will remain slow.

4.7 The Bad Control Problem through DAGs

Despite these limitations, there are areas where DAGs offer genuine insight that the PO framework does not provide as easily.

4.7.1 Clarity in the Bad Control Problem

The *bad control problem* is illustrated with remarkable clarity in the DAG framework. Simpson's Paradox, which we discussed in Chapter 1, is immediately transparent when drawn as a graph: the "paradox" disappears once the DAG tells you whether the third variable is a confounder, which must be conditioned on, or a collider, which must not be.

In the PO framework, the standard advice for selecting control variables is Angrist's rule of thumb: "only control for variables determined *before* the treatment." This is useful, but it is actually *wrong* in general, as the concept of M-bias shows.

4.7.2 M-bias

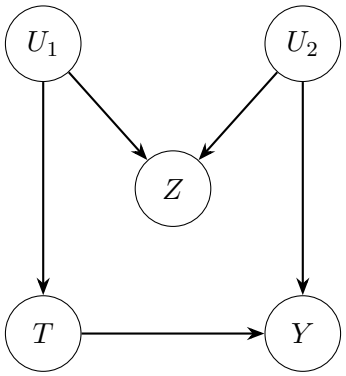


Figure 4.9: M-bias: Z is a pre-treatment variable, but it is a collider on the path $T \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$.

In this graph, Z is a *pre-treatment* variable. Angrist's rule of thumb would say: control for Z, since it is determined before T. But Z is a *collider* on the path $T \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$. Without conditioning on Z, this path is blocked (because Z is a collider). Conditioning on Z *opens* this path, creating a spurious association between T and Y.

The Lesson of M-bias

We should not control for colliders, and colliders do not necessarily occur *after* treatment. M-bias is the case where the collider is a pre-determined variable. The DAG framework provides a powerful and general tool for selecting control variables, given our assumptions about the causal structure. It forces us to state our causal assumptions explicitly and transparently.

4.8 Control Variable Issues: A Final Conclusion

Let us conclude this chapter by reviewing and discussing the issue of control variable selection. In earlier periods of empirical economics, economists often paid limited attention to the choice of control variables. Researchers frequently selected controls pretty arbitrarily, leaving substantial room for cherry-picking. As applied econometric methods have advanced, and as economists have increasingly incorporated insights from fields such as computer science and statistics, the discipline has become much more rigorous in its approach to control variable selection. Broadly speaking, there are two perspectives on how to select control variables. These perspectives reflect two different scientific philosophies and two broad approaches to empirical economic research: theory or economic knowledge driven approaches versus data driven approaches.

First, control variables can be selected based on the underlying causal structure. The key idea is to propose a causal graph that represents the data generating process we believe to be true, and then determine which variables are confounders and which are colliders. To identify the treatment effect of interest, we should control for confounders while avoiding conditioning on colliders. The procedure is inherently more subjective and relies heavily on economic theory and institutional knowledge. We include certain control variables because we believe they play an important role within the causal framework we have specified.

Second, control variables can be selected based on model fit. This relates closely to the model selection problem discussed in Chapter 3. Once the basic identification assumptions implied by the causal structure are satisfied, researchers may seek to improve the goodness of fit of the model, especially its out of sample predictive performance. This is where machine learning methods become particularly useful. Such methods provide automated procedures for selecting control variables that optimize predictive accuracy, thereby reducing researcher discretion and limiting the scope for cherry-picking. This approach is more objective and data driven, relying less on prior human judgment and more on patterns revealed by the data.

One final point is particularly important. As economists, we have strong economic theory foundations and our primary concern is causal identification. Therefore, control variables should first be selected based on the causal structure we propose, and only afterward should considerations of model fit be incorporated. For example, one should always control for education in a wage equation, even if a LASSO procedure suggests

otherwise.

4.9 Conclusion

The DAG approach to causality fully deserves the attention of all economists. It offers several genuine advantages:

- **Clarity:** DAGs force researchers to state their causal assumptions explicitly, making them transparent and debatable.
- **Control selection:** The backdoor criterion provides a principled, systematic method for choosing control variables, one that avoids pitfalls like M-bias that simpler rules of thumb miss.
- **Complex models:** For models with many variables, the do-calculus provides algorithmic tools for determining identifiability.
- **Novel strategies:** The frontdoor criterion demonstrates that identification strategies exist beyond what the PO framework naturally suggests.

However, the DAG approach also has significant limitations for economics:

- It requires an ex ante causal structure, with little guidance on where this structure comes from.
- It cannot easily express shape restrictions (monotonicity, convexity) or handle treatment effect heterogeneity.
- It cannot represent simultaneity and equilibrium, which are central to many economic models.
- It lacks concrete empirical examples demonstrating its value over existing methods.

As Imbens (2020) concludes, the PO and DAG frameworks are *complementary*, with different strengths for different questions. The honest assessment is that, for the kinds of questions applied economists typically ask (with relatively few variables, strong institutional knowledge, and a focus on heterogeneity and inference), the PO framework remains the more natural choice. But understanding DAGs makes you a better applied researcher: it sharpens your thinking about what you are controlling for and why.

How to bring the insights of DAGs more fully into applied economics remains a very

open question. There are opportunities here for the next generation of researchers.

Homework Problems

1. **Collider Bias Proof.**

   Consider the collider structure $X_1 \to X_2 \leftarrow X_3$.

   (a) Show that X_1 and X_3 are marginally independent: $P(x_1, x_3) = P(x_1) \cdot P(x_3)$.

   (b) Prove that conditioning on X_2 generally creates a dependence between X_1 and X_3. That is, show that $P(x_1, x_3 \mid x_2) \neq P(x_1 \mid x_2) \cdot P(x_3 \mid x_2)$ in general.

   Hint: Replicate the proof strategy used for chains and forks in Section 4.4.2.

2. **d-Separation Practice.**

   Consider the following DAG:

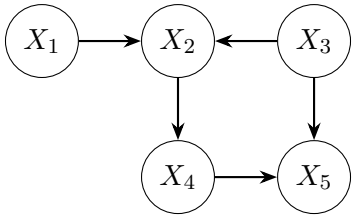


   Figure 4.10: DAG for the d-separation practice problem.

   For each of the following, determine whether the two variables are d-separated by the given conditioning set, and explain your reasoning:

   (a) X_1 and X_3, conditioning on $\emptyset$.

   (b) X_1 and X_3, conditioning on $\{X_2\}$.

   (c) X_1 and X_5, conditioning on $\{X_3\}$.

   (d) X_1 and X_5, conditioning on $\{X_2, X_3\}$.

3. **Backdoor Criterion Application.**

   Consider the following DAG representing the effect of a job training program (T) on wages (Y):

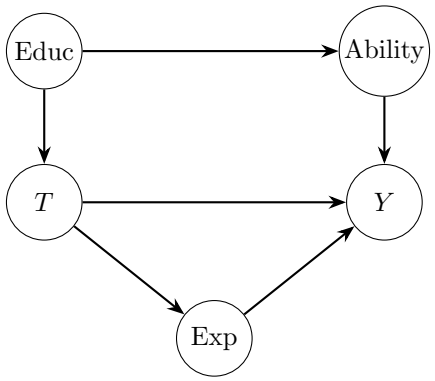


Figure 4.11: DAG for the backdoor criterion application problem.

Here Educ = education level, Ability = innate ability, Exp = work experience (post-training), T = training program, Y = wages.

(a) List all backdoor paths from T to Y.

(b) Does {Educ} satisfy the backdoor criterion? Why or why not?

(c) Does {Exp} satisfy the backdoor criterion? Why or why not?

(d) Find a minimal set that satisfies the backdoor criterion.

Practice Example: Backdoor Adjustment in a Simulated DAG

In this practice example, we simulate data from a known DAG and verify that the backdoor adjustment formula recovers the true causal effect, while a naive regression does not.

Data Generating Process. Consider the following DAG:

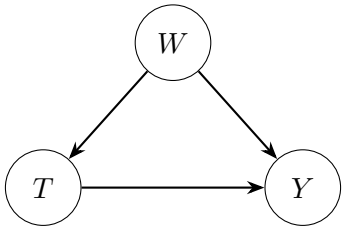


Figure 4.12: DAG for the practice example: W is the confounder of T and Y.

The structural equations are:

$$
\begin{aligned}
W_i &\sim \mathcal{N}(0,1), \\
T_i &= 0.8W_i + \varepsilon_{T,i}, \quad \varepsilon_{T,i} \sim \mathcal{N}(0,1), \\
Y_i &= \underbrace{2}_{\text{true causal effect}} \cdot T_i + 1.5W_i + \varepsilon_{Y,i}, \quad \varepsilon_{Y,i} \sim \mathcal{N}(0,1).
\end{aligned}
$$

The true causal effect of T on Y is $\beta = 2$. We generate $N = 2000$ observations.

Analysis. The Stata do-file (`src/chapter 4/ch4_dag.do`) performs the following steps:

1. **Simulate data** from the structural equations above.
2. **Naive regression:** Regress Y on T only (ignoring W). This produces a biased estimate because the backdoor path $T \leftarrow W \rightarrow Y$ is open.
3. **Backdoor adjustment:** Regress Y on T and W, controlling for the confounder. This blocks the backdoor path and recovers the true causal effect.
4. **Compare** the two estimates to the true value $\beta = 2$.

Expected results. The naive regression should overestimate the causal effect (because W positively affects both T and Y, creating upward omitted variable bias); the population value of the naive slope is $\text{Cov}(T,Y)/\text{Var}(T) = 4.48/1.64$, approximately 2.73. The controlled regression should produce an estimate close to 2.

Key lessons:

- The naive regression conflates the causal effect with the confounding association through W.
- The backdoor adjustment (controlling for W) isolates the causal channel $T \rightarrow Y$.
- This example illustrates the core logic of the backdoor criterion: identify all backdoor paths, find a valid conditioning set, and control for it.

The do-file is located at `src/chapter 4/ch4_dag.do` and the simulated dataset at `data/chapter 4/ch4_dag.dta`.

Part II

Instrumental Variables

Chapter 5

Introduction to Instrumental Variables and the Endogeneity Problem

5.1 Introduction: The Endogeneity Problem

In Chapter 1, we introduced the potential outcomes framework and emphasized that the central challenge of causal inference is the selection problem: individuals who receive treatment may differ systematically from those who do not, even in the absence of treatment. When we can run a randomized controlled trial (RCT), the treatment assignment is independent of potential outcomes, and a simple comparison of means identifies the average treatment effect. In practice, however, most economic variables of interest are not randomly assigned. This chapter introduces the most important tool economists have developed to address this problem: **instrumental variables (IV)**.

Consider one of the oldest and most studied questions in labor economics: what is the causal effect of schooling on wages? For individual i, suppose the relationship is linear and homogeneous (constant treatment effect):

$$Y_i = \alpha + \rho s_i + \eta_i, \tag{5.1}$$

where Y_i is wage, s_i is years of schooling, and η_i is an unobserved error term that captures everything else affecting wages. If schooling were randomly assigned, then s_i

would be independent of η_i, and the OLS estimator of ρ would consistently estimate the average treatment effect (ATE). But schooling is usually an endogenous choice made by the individual. People who attend college tend to have higher innate ability, stronger motivation, or better family backgrounds, all of which independently affect wages. This is **selection bias**.

To make the problem concrete, suppose that ability A_i is the key unobserved confounder. We can decompose the error as:

$$\eta_i = \gamma A_i + \nu_i, \tag{5.2}$$

where ν_i is independent of s_i. Plugging (5.2) into (5.1):

$$Y_i = \alpha + \rho s_i + \gamma A_i + \nu_i. \tag{5.3}$$

If A_i is observed, we can simply control for it and OLS on (5.3) consistently estimates ρ. But if A_i is unobserved, we have an **omitted variable bias** problem: OLS on (5.1) conflates the causal effect of schooling with the correlation between schooling and ability.

What can we do? This is where instrumental variables come in.

5.2 Simple IV: Definition and Identification

Let us begin with the simplest setting: a single endogenous variable, a single instrument, and a constant treatment effect. We return to the model in (5.1):

$$Y_i = \alpha + \rho s_i + \eta_i.$$

Definition 5.2.1 (Instrumental Variable)**.** A variable z_i is called an **instrumental variable** for s_i if it satisfies two conditions:

1. **Exogeneity (Exclusion Restriction):** $z_i \perp\!\!\!\perp \eta_i$. The instrument is independent of the unobserved determinants of the outcome.
2. **Relevance (Existence of First Stage):** $\mathrm{Cov}(s_i, z_i) \neq 0$. The instrument is correlated with the endogenous variable.

The exclusion restriction says that z_i affects Y_i *only through* s_i; there is no direct effect of z_i on Y_i other than through its influence on the endogenous variable. The relevance condition ensures that z_i actually shifts s_i; without it, we have no "lever" to move the endogenous variable.

5.2.1 Identification via Covariances

How does an instrument help us identify ρ? The key insight is to compute the covariance of z_i with both sides of (5.1):

$$\text{Cov}(z_i, Y_i) = \text{Cov}(z_i, \alpha + \rho s_i + \eta_i) = \rho \, \text{Cov}(z_i, s_i),$$

where the last equality uses the exogeneity condition: independence of z_i and η_i implies $\text{Cov}(z_i, \eta_i) = 0$. Solving for ρ:

IV Identification Formula

$$\rho = \frac{\text{Cov}(z_i, Y_i)}{\text{Cov}(z_i, s_i)} = \frac{\text{Cov}(z_i, Y_i)/\text{Var}(z_i)}{\text{Cov}(z_i, s_i)/\text{Var}(z_i)}. \tag{5.4}$$

The causal effect is identified by the ratio of two covariances or, equivalently, the ratio of two regression coefficients.

When the instrument z_i is binary, this formula simplifies to the **Wald estimator**:

$$\rho = \frac{E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0]}{E[s_i \mid z_i = 1] - E[s_i \mid z_i = 0]}. \tag{5.5}$$

The numerator is the effect of the instrument on the outcome (the "reduced form"), and the denominator is the effect of the instrument on the endogenous variable (the "first stage"). The ATE is the ratio of the two.

5.2.2 The Wald Estimator

To see the connection more formally, consider two regressions:

$$s_i = \pi_{10} + \pi_1 z_i + \xi_{1i} \qquad \text{(First Stage)}, \tag{5.6}$$

$$Y_i = \pi_{20} + \pi_2 z_i + \xi_{2i} \qquad \text{(Reduced Form)}. \tag{5.7}$$

Each of π_1 and π_2 is a simple regression coefficient, hence a ratio of covariance to variance. From (5.4):

$$\rho = \frac{\pi_2}{\pi_1}.$$

The sample analogue, the **Wald/IV estimator**, is:

$$\hat{\rho}_{\text{wald}} = \frac{\hat{\pi}_2^{\text{ols}}}{\hat{\pi}_1^{\text{ols}}}. \tag{5.8}$$

5.3 Two Stage Least Squares (2SLS)

A more general estimator for IV is **Two Stage Least Squares (2SLS)**. It is simple to implement and economically intuitive.

Consider the structural equation and first-stage equation:

$$Y_i = X_i'\alpha + \rho s_i + \eta_i, \tag{5.9}$$

$$s_i = X_i'\pi_{10} + \pi_{11} z_i + \xi_{1i}, \tag{5.10}$$

where X_i is a vector of exogenous control variables and z_i is the instrument.

Substitute (5.10) into (5.9):

$$\begin{aligned} Y_i &= X_i'\alpha + \rho(X_i'\pi_{10} + \pi_{11} z_i + \xi_{1i}) + \eta_i \\ &= X_i'\alpha + \rho(X_i'\pi_{10} + \pi_{11} z_i) + \xi_{2i}, \end{aligned} \tag{5.11}$$

where $\xi_{2i} = \rho\xi_{1i} + \eta_i$. The term $(X_i'\pi_{10} + \pi_{11} z_i)$ is the population projection of s_i onto z_i and X_i. Because $z_i \perp \eta_i$ and X_i is exogenous, this projected part is uncorrelated with ξ_{2i}. The endogeneity of s_i came from the component ξ_{1i} that is correlated with η_i; by replacing s_i with its predicted value from the first stage, we purge that problematic component.

The 2SLS procedure is exactly what its name suggests:

1. **First Stage:** Regress s_i on both z_i and X_i to obtain the predicted value:

$$\hat{s}_i = X_i'\hat{\pi}_{10} + \hat{\pi}_{11} z_i.$$

2. **Second Stage:** Regress Y_i on the predicted value $\hat{s}_i$ and X_i:

$$Y_i = X_i'\alpha + \rho\hat{s}_i + \xi_{2i}'.$$

   The coefficient on $\hat{s}_i$ is the 2SLS estimate of ρ.

Several important points deserve emphasis:

1. **Include the same controls in both stages.** The control variables X_i must appear in both the first-stage and second-stage regressions.
2. **Never compute 2SLS by hand.** In Stata, use `ivregress 2sls` or `ivreg2`. If you manually run two OLS regressions, the second-stage standard errors will be wrong because they fail to account for the estimation uncertainty in $\hat{s}_i$.
3. **The first stage need not have a causal interpretation.** You can always run a regression without claiming causality. What matters is that the instrument predicts the endogenous variable. That said, in practice it is better if you have a reason to believe that z actually affects s; a strong theoretical story makes the first stage more credible.

There is a close connection between Wald and 2SLS estimators.

- The Wald estimator is only available when the number of endogenous variables equals the number of instruments (exactly identified).
- When the model is just-identified, the 2SLS estimator is algebraically identical to the Wald estimator.
- In general, 2SLS is more efficient, and it is the most efficient IV estimator under homoskedasticity.

5.4 IV with Heterogeneous Treatment Effects: LATE

So far, we have maintained the assumption of a constant treatment effect: every individual has the same causal effect ρ. In many applications, this is unrealistic. Some people benefit more from schooling than others; some patients respond more to a drug than others. What does the IV estimator identify when treatment effects are heterogeneous?

This question was answered in a landmark paper by Imbens and Angrist (1994). The answer is the **Local Average Treatment Effect (LATE)**.

5.4.1 Setting: The Vietnam Draft Lottery

To motivate the framework, consider the classic study by Angrist (1990) on the effect of military service on earnings. During the Vietnam War, young men in the United States were drafted into the army based on a lottery: each birthday was assigned a random draft lottery number, and men with numbers below a cutoff were likely to be drafted. The key variables are:

- Y_i: wage earnings,
- D_i: whether individual i served in the military (treatment),
- z_i: whether the draft lottery number is below the cutoff (instrument).

The draft number is a plausible instrument: it was randomly assigned (exogenous) and it affected the probability of military service (relevant), but it should not affect wages through any channel other than military service (exclusion restriction).

5.4.2 Potential Outcomes and Four Assumptions

We now define potential outcomes more carefully. Let $Y_i(d, z)$ denote the potential outcome (wage) for individual i given treatment status d and instrument value z. Let D_{1i} and D_{0i} denote the potential treatment status when the instrument takes value 1 and 0, respectively.

The LATE theorem requires four assumptions:

Assumption 5.4.1 (LATE). 1. **Independence:**

$$\{Y_i(D_{1i}, 1),\ Y_i(D_{0i}, 0),\ D_{1i},\ D_{0i}\} \perp\!\!\!\perp z_i.$$

The instrument is as good as randomly assigned: it is independent of both potential outcomes and potential treatment status (agent types).

2. **Exclusion Restriction:**

$$Y_i(d, 0) = Y_i(d, 1) \equiv Y_{di} \quad \text{for } d = 0, 1.$$

The instrument can affect the outcome *only* through treatment. In the draft example, the lottery number affects future wages only by changing military service, not through any other channel (e.g., education decisions).

3. **Existence of First Stage:**

$$E[D_{1i} - D_{0i}] \neq 0.$$

The instrument must shift the probability of treatment on average.

4. **Monotonicity:**

$$\forall\, i, \quad D_{1i} - D_{0i} \geq 0 \quad \text{(or vice versa)}.$$

The instrument moves treatment in the same direction for everyone (or has no effect). There are no **defiers**, individuals who do the opposite of what the instrument encourages.

5.4.3 Compliance Types

The monotonicity assumption partitions the population into three types based on how individuals respond to the instrument:

- **Compliers** ($D_{1i} > D_{0i}$): People who change their treatment status because of the instrument. In the draft example, these are men who served in the military only because their lottery number was below the cutoff.
- **Always-takers** ($D_{1i} = D_{0i} = 1$): People who always take treatment regardless of the instrument. These are men who would serve voluntarily even without being drafted.
- **Never-takers** ($D_{1i} = D_{0i} = 0$): People who never take treatment regardless of the instrument. These are men who would not serve even if drafted (e.g., obtained exemptions).

Monotonicity rules out a fourth type, **defiers** ($D_{1i} < D_{0i}$), individuals who would serve when their number is above the cutoff but refuse when it is below. This assumption is often plausible in practice.

5.4.4 The LATE Theorem

Theorem 5.4.1 (LATE Theorem, Theorem 4.4.1 in Angrist and Pischke, 2009)**.** *Under Assumptions 1–4, the IV (Wald) estimand identifies the average treatment effect for the complier subpopulation:*

$$\frac{E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0]}{E[D_i \mid z_i = 1] - E[D_i \mid z_i = 0]} = E[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}]. \tag{5.12}$$

Proof. By monotonicity there are no defiers, so the population is partitioned into three types: always-takers (A), compliers (C), and never-takers (N). Decompose the intention-to-treat (ITT) effect:

$$\begin{aligned}
&E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0] \\
&= P(A_i \mid z_i = 1)\, E[Y_{1i} \mid A_i, z_i = 1] + P(C_i \mid z_i = 1)\, E[Y_{1i} \mid C_i, z_i = 1] \\
&\quad + P(N_i \mid z_i = 1)\, E[Y_{0i} \mid N_i, z_i = 1] \\
&\quad - \big[P(A_i \mid z_i = 0)\, E[Y_{1i} \mid A_i, z_i = 0] + P(C_i \mid z_i = 0)\, E[Y_{0i} \mid C_i, z_i = 0] \\
&\quad + P(N_i \mid z_i = 0)\, E[Y_{0i} \mid N_i, z_i = 0]\big].
\end{aligned}$$

By the independence assumption, z_i is independent of compliance type and potential outcomes. Therefore:

- $P(A_i \mid z_i = 1) = P(A_i \mid z_i = 0) = P(A_i)$, and similarly for C and N.
- $E[Y_{di} \mid \text{type}, z_i = 1] = E[Y_{di} \mid \text{type}, z_i = 0]$ for each type.

The always-taker terms cancel (both equal $P(A_i)\, E[Y_{1i} \mid A_i]$). The never-taker terms cancel (both equal $P(N_i)\, E[Y_{0i} \mid N_i]$). Only the complier terms remain:

$$E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0] = P(C_i)\big(E[Y_{1i} \mid C_i] - E[Y_{0i} \mid C_i]\big).$$

Similarly, since only compliers change their treatment status with z_i, the denominator equals:

$$E[D_i \mid z_i = 1] - E[D_i \mid z_i = 0] = P(C_i).$$

Dividing:

$$\frac{E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0]}{E[D_i \mid z_i = 1] - E[D_i \mid z_i = 0]} = E[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}].$$

□

5.4.5 Interpreting LATE

LATE has several important properties:

- **LATE is internally valid:** it identifies a well-defined causal effect for a well-defined subpopulation (compliers).
- **LATE is policy-relevant:** compliers are precisely those whose behavior can be changed by the policy instrument. If a policymaker is considering expanding the draft, the LATE tells them the average effect on the marginal group that would be affected.
- **Monotonicity is crucial:** without it, the effects of compliers and defiers would be mixed together, and the IV estimand would not have a clear causal interpretation.

However, LATE also has important limitations:

- **External validity:** LATE is specific to the complier group induced by a particular instrument. If the policy changes, the complier group changes, and the LATE may differ.
- **Multi-valued instruments and treatments:** When the instrument or treatment is not binary, the number of compliance types grows exponentially, making the traditional LATE framework difficult to apply. As Pinto (2015) discusses, we need new approaches, such as combining IV with choice models, to handle these cases. This will be the subject of Chapter 6.

5.5 Multiple IV and the GMM Framework

So far, we have considered settings with a single endogenous variable and a single instrument. In practice, researchers often have multiple endogenous variables and multiple instruments. The **Generalized Method of Moments (GMM)** provides a unified framework for understanding all IV-related estimators.

5.5.1 Moment Equation Models

Let $g_i(\beta)$ be a known $l \times 1$ function of a $k \times 1$ parameter vector β.

Definition 5.5.1 (Moment Equation Model)**.** A moment equation model is defined by

the condition:

$$E[g_i(\beta)] = 0. \tag{5.13}$$

In this system, we have l known equations and k unknown parameters.

Example 5.5.1. The linear regression model is a moment equation model with $l = k$ and $g_i(\beta) = x_i(Y_i - x_i'\beta)$. The moment condition $E[x_i(Y_i - x_i'\beta)] = 0$ is just the population orthogonality condition for OLS.

Given that these l equations are not linearly dependent, the relationship between l and k determines the identification status:

- If $l = k$: the model is **just-identified**. There are exactly as many equations as unknowns.
- If $l > k$: the model is **over-identified**. There are more equations than unknowns.
- If $l < k$: the model is **under-identified**. There are fewer equations than unknowns, and β cannot be uniquely recovered.

5.5.2 Method of Moments and GMM

When $l = k$ (just-identified), we can solve the system of equations directly using the **Method of Moments Estimator (MME)**: set the sample analogue of the moment conditions to zero and solve for $\hat{\beta}$:

$$\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta}) = 0. \tag{5.14}$$

OLS is a special case: $\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} x_i(Y_i - x_i'\hat{\beta}) = 0$ gives exactly the OLS normal equations.

When $l > k$ (over-identified), we have more equations than unknowns, so in finite samples we generally cannot set all moment conditions to zero simultaneously. Here is an important conceptual difference we need to clarify. Although we assume that the moment conditions $E[g_i(\beta)] = 0$ hold in the population, their sample analogue does not hold exactly in finite samples because of sampling uncertainty.

Instead, we minimize the *distance* between the sample moment vector and zero:

$$J(\beta) = n\,\bar{g}_n(\beta)'\,W\,\bar{g}_n(\beta), \tag{5.15}$$

where W is a positive-definite **weighting matrix**. The **GMM estimator** is:

$$\hat{\beta}_{\text{gmm}} = \arg\min_{\beta} J(\beta). \tag{5.16}$$

The function J measures the weighted Euclidean distance between $\bar{g}_n$ and zero. The MME (and thus OLS) is a special case of GMM when $l = k$.

5.5.3 Linear GMM

In the linear IV model, let X_i denote the (potentially endogenous) regressors and Z_i the instruments. The moment conditions are:

$$E[g_i(\beta)] = E[Z_i(Y_i - X_i'\beta)] = 0. \tag{5.17}$$

Stacking over the sample, the GMM estimator minimizes:

$$\hat{\beta}_{\text{gmm}} = \arg\min_{\beta} \underbrace{\tfrac{1}{n}\left(Z'Y - Z'X\beta\right)' W \left(Z'Y - Z'X\beta\right)}_{J(\beta)}.$$

Here the dimension for X is $n \times k$, for Z is $n \times l$, for Y is $n \times 1$.

Theorem 5.5.1 (Linear GMM Estimator, Theorem 13.1 in Hansen, 2022)**.** *For the linear IV model with k (possibly endogenous) regressors and $l \geq k$ instruments:*

$$\hat{\beta}_{gmm} = (X'ZWZ'X)^{-1}(X'ZWZ'Y). \tag{5.18}$$

This is a remarkably general formula. Many familiar estimators are special cases, depending on the choice of W and the relationship between X and Z.

5.5.4 Special Cases of the Linear GMM Estimator

Case 1: OLS. When $X = Z$ (no endogenous variables) and $W = \mathbf{I}$:

$$\begin{aligned}
\hat{\beta}_{\text{gmm}} &= (X'X\mathbf{I}X'X)^{-1}(X'X\mathbf{I}X'Y) \\
&= (X'X)^{-1}(X'X)^{-1}(X'X)(X'Y) \\
&= (X'X)^{-1}X'Y = \hat{\beta}_{\text{ols}}.
\end{aligned}$$

The second line follows because $X'X$ is a square matrix when $X = Z$ (so $l = k$).

Case 2: Wald/IV Estimator. When $l = k$ (just-identified) and $W = \mathbf{I}$:

$$\begin{aligned}\hat{\beta}_{\text{gmm}} &= (X'Z\mathbf{I}Z'X)^{-1}(X'Z\mathbf{I}Z'Y) \\ &= (Z'X)^{-1}(X'Z)^{-1}(X'Z)(Z'Y) \\ &= (Z'X)^{-1}Z'Y.\end{aligned}$$

Since $l = k$, the matrix $X'Z$ is square and invertible. This is the Wald/IV estimator.

Case 3: 2SLS Estimator. Let $P_z = Z(Z'Z)^{-1}Z'$ be the projection matrix onto the column space of Z, so that the first-stage fitted values are $\hat{X} = P_z X$. Projection matrix P has an important property called *idempotent*, such that $P \cdot P = P$.

When $W = (Z'Z)^{-1}$:

$$\begin{aligned}\hat{\beta}_{\text{gmm}} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'Y) \\ &= (X'P_zX)^{-1}X'P_zY \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y.\end{aligned}$$

This is exactly the 2SLS estimator: regress Y on $\hat{X}$, where $\hat{X}$ is the first-stage prediction.

Remark 5.5.1. The fact that OLS, Wald, and 2SLS are all special cases of GMM is not merely an algebraic curiosity. It means that the asymptotic theory for GMM (consistency, asymptotic normality, efficient weighting) applies to all of these estimators as special cases. Understanding GMM gives you a unified toolkit for inference in IV models.

5.5.5 Over-identification Test

When we have more instruments than endogenous variables ($l > k$), we can test whether the moment conditions hold, that is, whether the instruments are valid. The logic is intuitive: if all instruments are valid, the minimized value of J should be close to zero. The test result below builds on Hansen (2022).

Theorem 5.5.2 (Hansen's J-Test, Theorem 13.14 in Hansen, 2022)**.** *Under the null hy-*

pothesis that all moment conditions are satisfied, when J is evaluated with the efficient weight matrix $W = \left(E[g_i g_i']\right)^{-1}$ (or a consistent estimate of it), as $n \to \infty$:

$$J = J(\hat{\beta}_{gmm}) \xrightarrow{d} \chi^2_{l-k}. \tag{5.19}$$

For c satisfying $\alpha = 1 - G_{l-k}(c)$, where G_{l-k} denotes the χ^2_{l-k} distribution function, we reject H_0 if $J > c$. The test has asymptotic size α.

Caution with the Over-identification Test

The J-test has several important caveats:

1. **You should *hope* for non-rejection.** A valid IV strategy implies that J should be small. Rejection means something is wrong with at least one of the moment conditions.
2. **Feasibility requires over-identification.** The test can only be conducted when $l > k$. In the just-identified case, $J = 0$ by construction, and there is nothing to test.
3. **Rejection has multiple interpretations.** The J-test is a specification test: rejection does not necessarily mean $E[g_i] \neq 0$. It could also reflect non-linearity or other forms of model misspecification.
4. **Non-rejection does not prove validity.** The J-test has power against specific alternatives, but it cannot detect all forms of instrument invalidity.
5. **Finite-sample size distortion.** The actual rejection rate in finite samples tends to be too large (the test over-rejects).

5.6 Oster Bound: Endogeneity without IV

Finding a valid instrument is often extremely difficult. In many empirical settings, no credible instrument is available. What can we do when we face an endogeneity problem but cannot find an IV?

Oster (2019) proposes a method, **Oster bound**, that uses the pattern of coefficient stability and changes in R^2 across specifications to assess the robustness of OLS estimates to omitted variable bias. It is important to be clear about what this method does and does not do.

5.6.1 Point Identification vs. Set Identification

Before presenting the method, it is useful to clarify two concepts:

Definition 5.6.1 (Point Identification)**.** A parameter β is **point-identified** if there is a unique mapping from the data distribution to the parameter value. No other parameter value could generate the same data.

Definition 5.6.2 (Set Identification)**.** A parameter β is **set-identified** if the data restrict β to lie within a set, but do not uniquely pin it down. No parameter value *outside* the identified set is consistent with the data.

IV provides point identification of ρ (under its assumptions). The Oster bound provides only *set identification*: it gives a range within which the true effect plausibly lies. It is not a method for point estimation. It provides suggestive evidence when nothing else is available.

5.6.2 The Intuition

The core intuition of the Oster bound is simple and elegant: **the relationship between the treatment variable and unobservable confounders can be partially recovered from the relationship between the treatment variable and observable controls.**

Consider the following thought experiment. You run two regressions:

1. A "short" regression of Y on X alone (no controls), obtaining coefficient $\mathring{\beta}$ and R-squared $\mathring{R}$.
2. An "intermediate" regression of Y on X and a set of observed controls ω, obtaining $\tilde{\beta}$ and $\tilde{R}$.

If there is large omitted variable bias, then when we add informative controls, the coefficient on X should change substantially and the R^2 should increase a lot. The Oster bound exploits this logic: if the coefficient is *stable* (changes only a little) when we add a *strong, informative* control (one that substantially increases R^2), this is evidence that omitted variable bias is small.

5.6.3 The Formal Setup

Suppose we are interested in the effect of X on Y. There are two sets of variables W_1 and W_2, both correlated with X and Y. The variable W_1 can be represented by some observed proxies ω, while W_2 is unobservable. The model is:

$$\begin{aligned} Y &= \beta X + \Psi\omega + W_2 + \epsilon, \\ W_1 &= \Psi\omega, \end{aligned} \tag{5.20}$$

where W_1 and W_2 are assumed to be orthogonal to each other.

Define the **proportional selection relationship** δ:

$$\delta \frac{\sigma_{1X}}{\sigma_1^2} = \frac{\sigma_{2X}}{\sigma_2^2}, \tag{5.21}$$

where $\sigma_{jX} = \text{Cov}(W_j, X)$ and $\sigma_j^2 = \text{Var}(W_j)$ for $j = 1, 2$. The parameter δ measures the relative importance of the observed controls W_1 versus the unobserved confounders W_2 in their relationship with the treatment X:

- When δ is large, the unobserved variable W_2 is relatively more important than the observed control W_1.
- When $\delta = 1$, the unobservable and observable are equally related to the treatment.

5.6.4 Three Regressions, Two Key Parameters

Denote the coefficient on X and the R-squared from three regressions:

1. **Short regression:** Y on $X \Rightarrow \mathring{\beta}$, $\mathring{R}$.
2. **Intermediate regression:** Y on X and $\omega \Rightarrow \tilde{\beta}$, $\tilde{R}$.
3. **Full regression:** Y on X, ω, and $W_2 \Rightarrow R_{\max}$.

We can compute $\mathring{\beta}$, $\mathring{R}$, $\tilde{\beta}$, and $\tilde{R}$ from the data. The two unknowns are δ (the proportional selection parameter) and $R_{\max}$ (the R-squared from the hypothetical regression that includes all confounders). Given $\mathring{R}$ and $\tilde{R}$, we can infer how much variation is explained by observed controls; $R_{\max}$ tells us the share explained by the additional unobserved control W_2.

5.6.5 Key Theoretical Results

Oster (2019) provides three propositions connecting δ, $R_{\max}$, and bias.

Proposition 5.6.1 (Proposition 1 in Oster (2019): Debiased Estimator under $\delta = 1$)**.** *When $\delta = 1$ (observables and unobservables are equally important), the bias-adjusted estimator is*

$$\beta^* = \tilde{\beta} - \underbrace{[\mathring{\beta} - \tilde{\beta}]\frac{R_{\max} - \tilde{R}}{\tilde{R} - \mathring{R}}}_{bias}, \tag{5.22}$$

and this debiased estimator is asymptotically consistent: $\beta^* \xrightarrow{p} \beta$.

The formula has an intuitive interpretation. The bias is positively related to the *coefficient change* $\mathring{\beta} - \tilde{\beta}$ (how much the coefficient moves when controls are added) and the ratio $\frac{R_{\max} - \tilde{R}}{\tilde{R} - \mathring{R}}$ (how much additional variation the unobservables explain, relative to what the observables explained).

Proposition 5.6.2 (Proposition 2 in Oster (2019))**.** *Given δ and $R_{\max}$, we can calculate the bias and find a debiased estimator. In some cases, there may be multiple solutions, requiring solution selection. In short: $\delta, R_{\max} \to bias, \beta$.*

Proposition 5.6.3 (Proposition 3 in Oster (2019))**.** *Given $R_{\max}$ and any targeted value of the treatment effect β, we can find the value of δ that makes the bias zero. In short: $R_{\max}, \beta, bias = 0 \to \delta$.*

Of course, we never observe δ or $R_{\max}$ (since W_2 is unobservable). But these propositions give us a powerful tool for *bounding* β under assumptions about these parameters.

5.6.6 Implementation in Practice

There are two practical methods for using the Oster bound:

Method 1: How robust is our result? Assume a value for $R_{\max}$ and calculate the value of δ for which $\beta = 0$ based on Proposition 5.6.3. This tells us how important unobservables would need to be (relative to observables) to completely erase the estimated effect.

1. As a rule of thumb, set $R_{\max} = \min\{1,\ 1.3\tilde{R}\}$. The factor 1.3 is derived from letting 90% of RCT studies in top journals pass this test.

2. Set $\beta = 0$ and find the corresponding δ.
3. If $\delta > 1$, our results are robust: unobservables would need to be more important than observables to overturn the finding.

Method 2: What is a conservative bound? Assume conservative values for both $R_{\max}$ and δ, and calculate the debiased estimate β^* based on Proposition 5.6.2.

1. Set $R_{\max} = \min\{1,\ 1.3\tilde{R}\}$ and $\delta = 1$.
2. Calculate β^* using (5.22).
3. The conservative bound on the treatment effect is $[\tilde{\beta}, \beta^*]$.
4. If the interval does not contain zero, the result is robust.

Limitations of the Oster Bound

1. The Oster bound is a **last resort**: use it when no valid instrument or other identification strategy is available.
2. The choice of $R_{\max}$ and δ is inherently arbitrary. The bound provides a *sense* of robustness, not a definitive answer.
3. It can also be used as a robustness check alongside other methods.
4. An application in economics: Clark et al. (2021) use the Oster bound to assess the robustness of their estimates of the effect of online learning on student performance during the COVID-19 pandemic.

5.7 Conclusion

This chapter introduced the instrumental variables framework as the primary tool for addressing endogeneity in empirical economics. The key takeaways are:

1. **IV requires two assumptions:** the exclusion restriction (the instrument affects the outcome only through the endogenous variable) and the existence of a first stage (the instrument is correlated with the endogenous variable). Both must be carefully justified in any application.
2. **With heterogeneous treatment effects and a single binary endogenous variable and instrument,** the IV/Wald estimator identifies the **Local Average Treatment Effect (LATE)**, the average effect for the complier subpopula-

tion. LATE is internally valid and policy-relevant, but not necessarily externally valid.

3. **GMM is the general framework** for IV estimation. OLS, the Wald estimator, and 2SLS are all special cases of the linear GMM estimator, differing only in the weighting matrix and the relationship between the number of regressors and the number of instruments.

4. **Over-identification tests** (Hansen's J-test) allow us to assess whether the moment conditions hold, but they should be interpreted cautiously.

5. **When no valid IV is available,** the Oster bound provides a way to assess robustness to omitted variable bias. It is based on the idea that the relationship between treatment and unobservables can be inferred from the relationship between treatment and observables. However, it provides only set identification and relies on somewhat arbitrary parameter choices.

In the next chapter, we will relax the remaining simplifying assumptions: we will move beyond binary variables and binary instruments, and consider how to combine IV with choice models to handle more complex settings.

5.8 Application of IV: Angrist (1990)

We have invoked the Vietnam draft lottery several times in this chapter as the canonical motivating example for IV with heterogeneous treatment effects. The empirical study that established this design as a textbook reference is Angrist (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," published in the *American Economic Review*. The research question is simple but consequential: did military service in the Vietnam era cause a long term reduction in civilian earnings? The naive answer, comparing earnings of veterans and nonveterans, is contaminated by selection into the military: men who enlisted or were drafted differ from those who did not in many ways that also affect earnings (education, health, family background, expected labor market prospects). Without a credible source of exogenous variation, the OLS comparison cannot be given a causal interpretation.

The methodological contribution of the paper is to use the Vietnam era draft lottery as an instrument for veteran status. From 1970 to 1972, the United States Selective

Service assigned each draft age man a random sequence number based on his date of birth. Men whose number fell below an administratively determined ceiling were called for induction, while those above the ceiling were exempt. The lottery number is plausibly random with respect to potential earnings and affects civilian earnings only through its effect on the probability of military service, so it satisfies the relevance, exclusion, and independence conditions discussed in Sections 5.2 and 5.4. Angrist constructs draft eligibility as the binary instrument and uses it together with Social Security administrative earnings records, computing Wald style estimates by birth cohort and confirming the results with 2SLS that absorbs cohort fixed effects. Under the heterogeneous treatment effects framework of Section 5.4, the estimand is a LATE: the average earnings loss for compliers, that is, men who served because they were draft eligible but would not have served otherwise.

The main empirical finding is that Vietnam era military service caused a sizable and persistent decline in the earnings of white veterans, on the order of fifteen percent of nonveteran earnings in the early 1980s, roughly a decade after discharge. For nonwhite men the estimated effect is much smaller and statistically indistinguishable from zero. Beyond the substantive finding, the paper has been hugely influential as a methodological template: it shows in a single setting how to combine a randomized policy lever with administrative data and the IV/LATE machinery to recover a credible causal effect. I strongly encourage students to read the original article to see how the formal framework of this chapter plays out in a real empirical study.

Homework Problems

1. **IV Identification with Covariates.**

   Consider the model

   $$Y_i = \alpha + \beta_1 X_i + \beta_2 D_i + \epsilon_i,$$

   where D_i is endogenous ($\text{Cov}(D_i, \epsilon_i) \neq 0$) and X_i is exogenous. Suppose z_i is a valid instrument: $\text{Cov}(z_i, \epsilon_i) = 0$ and $\text{Cov}(z_i, D_i) \neq 0$.

   (a) Show that the IV estimator of β_2 can be written as a ratio of covariances involving the residuals from projecting Y_i and D_i on X_i.

   (b) Explain intuitively why we need to "partial out" X_i before applying IV.

2. **LATE with Compliance Data.**

   Suppose 1000 individuals are randomly assigned to treatment ($z_i = 1$, 500 people) or control ($z_i = 0$, 500 people). Among those assigned to treatment, 300 actually take treatment ($D_i = 1$) and 200 do not. Among those assigned to control, 50 take treatment and 450 do not. Average outcomes: $\bar{Y}(z = 1) = 10$, $\bar{Y}(z = 0) = 8$.

   (a) Calculate the Wald/LATE estimator.

   (b) What fraction of the population are compliers? Always-takers? Never-takers?

   (c) Under what assumption can we rule out defiers?

3. **Oster Bound Calculation.**

   A researcher estimates the effect of class size on test scores. The short regression (no controls) gives $\mathring{\beta} = -2.5$ with $\mathring{R} = 0.04$. Adding school-level controls gives $\tilde{\beta} = -1.8$ with $\tilde{R} = 0.22$.

   (a) Using $R_{\max} = 1.3 \times \tilde{R}$ and $\delta = 1$, compute the bias-adjusted estimate β^*.

   (b) Does the interval $[\tilde{\beta}, \beta^*]$ contain zero? What does this tell us about robustness?

   (c) Using the same $R_{\max}$, find the value of δ that would make $\beta^* = 0$. Is the result robust?

Practice Example: Returns to Schooling

In this practice example, we simulate data to illustrate the endogeneity problem and the IV solution. The setting mirrors the classic returns-to-schooling question.

Data Generating Process. We simulate $N = 2{,}000$ individuals. Each individual has an unobserved ability $A_i \sim N(0, 1)$. A binary instrument z_i (e.g., proximity to a college) is randomly assigned with $P(z_i = 1) = 0.5$, independent of ability. Schooling is determined by:

$$s_i = 12 + 2z_i + 1.5A_i + u_i, \quad u_i \sim N(0, 1),$$

so the instrument shifts schooling by 2 years on average. Wages are:

$$Y_i = 20 + 3s_i + 2A_i + \epsilon_i, \quad \epsilon_i \sim N(0, 25),$$

so the true causal effect of one additional year of schooling is $\rho = 3$. Because ability A_i enters both equations, OLS on the short regression will be upward biased.

Analysis. The Stata do-file (located at `src/chapter 5/ch5_iv.do`) performs the following steps:

1. Simulate the data and save as `data/chapter 5/ch5_iv.dta`.
2. Run OLS of Y on s (biased estimate).
3. Run the first-stage regression of s on z.
4. Run the reduced-form regression of Y on z.
5. Compute the Wald estimator as the ratio of reduced-form to first-stage coefficients.
6. Run 2SLS using `ivregress 2sls`.
7. Run the manual two step procedure (OLS of Y on the first-stage fitted values $\hat{s}$) and compare its standard error with the one reported by `ivregress`.
8. Compare OLS and IV estimates to the true effect $\rho = 3$.

Expected Results. The OLS estimate should be substantially above 3 (upward biased due to ability). The IV/2SLS estimate should be close to 3. The first-stage F-statistic should be large, confirming instrument relevance. The manual two step procedure reproduces the 2SLS point estimate exactly but reports a different standard error, which illustrates why 2SLS should never be computed by hand.

Chapter 6

IV beyond LATE

6.1 Introduction: The Limits of LATE

In the previous chapter we developed the Local Average Treatment Effect interpretation of IV, following Imbens and Angrist (1994). LATE is elegant, transparent, and tightly linked to a research design: if the instrument is randomly assigned, satisfies exclusion, and induces a monotone change in treatment, then 2SLS recovers the average causal effect for the subpopulation of compliers. Almost every modern IV paper in applied economics is, implicitly or explicitly, an exercise in interpreting an estimate as a LATE.

This is a real achievement, but LATE has two important limitations that the original framework cannot easily overcome (Heckman and Vytlacil, 2007a,b).

First, the standard LATE theorem assumes a single binary instrument and a single binary treatment. In practice we often have multiple instruments, multivalued instruments, or a multivalued treatment such as years of schooling. Once we leave the clean 2×2 world, the IV/2SLS estimand becomes a weighted average of treatment effects, where the weights depend on the first stage, the structure of the instrument, and the behavior of agents. Understanding what 2SLS is actually estimating in these settings requires more work, and in some cases requires us to import economic theory into the identification argument.

Second, LATE is internally valid but not externally valid. The complier subpopulation is defined ex post, by the way agents respond to a particular instrument. If the policy environment changes, the compliers change too. The Vietnam draft compliers

of Angrist (1990) are not the same group as the Vietnam volunteers, and they are not the same group as those induced to enlist by a future policy. We can use LATE to evaluate the policy that produced our instrument, but we cannot generally use it to forecast the effect of a new policy on a different group of marginal participants.

This chapter does two things to address these limitations.

Part one generalizes the LATE interpretation within its original design-based framework. We work through three cases in turn: multiple binary instruments, a multivalued (ordered) treatment with a binary instrument, and a multivalued instrument. The first two have clean weighted-average interpretations that follow from imposing monotonicity. The third reveals that monotonicity is not innocuous when the instrument takes more than two values, and forces us to think more carefully about the choice problem agents are solving. We introduce the weak and strong axioms of revealed preference (WARP and SARP) and walk through the Pinto (2015) Moving to Opportunity (MTO) example to see how economic theory regularizes IV identification when the design alone is not enough.

Part two introduces the Marginal Treatment Effect (MTE) framework of Heckman and Vytlacil (2005, 2007a,b). The MTE is the average treatment effect for agents who are indifferent between taking and not taking treatment, indexed by their observable covariates X and a one-dimensional unobserved resistance U_D. It is a deep structural parameter that does not depend on which instrument we use; LATE, ATE, ATT, ATUT, and the IV and OLS estimands are all weighted averages of the MTE with different weights. We define the MTE and show how to estimate it using the Local Instrumental Variable (LIV) estimand.

We close the chapter with the Kline and Walters (2016) application to Head Start, which illustrates how an IV strategy can be interpreted from a more general behavioral model perspective and taken beyond any single LATE estimate.

The unifying theme of this chapter is that *treatment selection is intrinsically part of any IV problem.* The pure design-based approach hides the selection problem inside the monotonicity assumption. When we let the choice structure come back into the open, IV becomes more powerful, more informative, and more transparently linked to the economic primitives we usually care about.

6.2 Choice Model Embedded in IV

Before we generalize anything, it is worth recognizing that the LATE framework already contains a choice model. We just used it without naming it.

Recall the LATE theorem from Chapter 5. With binary IV $Z \in \{0, 1\}$ and binary treatment $D \in \{0, 1\}$, agents fall into four behavior types based on the pair $(D(0), D(1))$:

- Always-takers (A): $D(0) = D(1) = 1$.
- Never-takers (N): $D(0) = D(1) = 0$.
- Compliers (C): $D(0) = 0$, $D(1) = 1$.
- Defiers (Def): $D(0) = 1$, $D(1) = 0$.

The monotonicity assumption rules out defiers. Where does this assumption come from? It is not a statistical convenience. It comes from a model of choice. If the instrument Z shifts the cost of treatment, and treatment is a non-Giffen good, then a lower cost can only weakly increase the probability of taking treatment. Formally: if an agent chooses $D = 1$ when the price is high ($Z = 0$), then he or she will also choose $D = 1$ when the price is lower ($Z = 1$). Defiers are ruled out by the law of demand.

This is more than a curiosity. It tells us that whenever we invoke monotonicity, we are implicitly invoking an economic model of how agents respond to incentives. In the 2×2 case the model is so simple it disappears from view. When we move to multivalued instruments and treatments, the underlying choice model has to do more work, and the assumptions we need become richer and more substantive. The slogan of this chapter is exactly that: when statistical tools run out, economics is what you have left.

With this perspective in mind, let us see what happens as we move to more general IV settings.

6.3 Generalizing LATE: Three Cases

6.3.1 Recap: Binary Instrument, Binary Treatment

In the 2×2 case there are four observed conditional expectations, one for each combination of Z and D:

$$
\begin{aligned}
E[Y \mid Z=1, D=1] &= P(A \mid Z=1, D=1)\, E[Y_1 \mid Z=1, D=1, A] \\
&\quad + P(C \mid Z=1, D=1)\, E[Y_1 \mid Z=1, D=1, C], \\
E[Y \mid Z=1, D=0] &= E[Y_0 \mid Z=1, D=0, N], \\
E[Y \mid Z=0, D=1] &= E[Y_1 \mid Z=0, D=1, A], \\
E[Y \mid Z=0, D=0] &= P(N \mid Z=0, D=0)\, E[Y_0 \mid Z=0, D=0, N] \\
&\quad + P(C \mid Z=0, D=0)\, E[Y_0 \mid Z=0, D=0, C].
\end{aligned}
$$

The left hand side is observed; the right hand side mixes potential outcomes that we cannot observe directly. Under monotonicity (no defiers) and the independence of Z, this system of four equations can be inverted to identify LATE, the average treatment effect for compliers. The tree of behavior types underlying the system is shown in Figure 6.1.

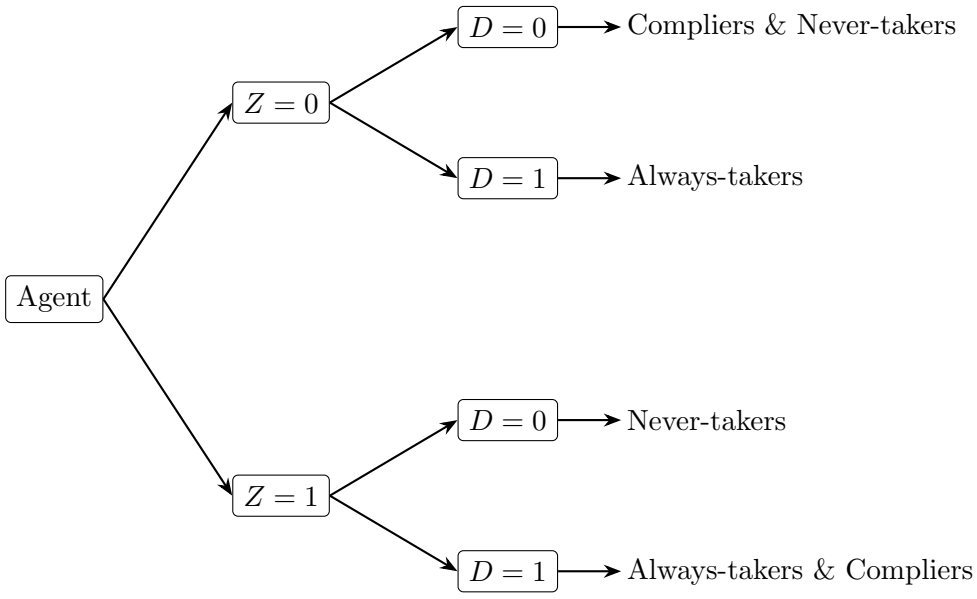


Figure 6.1: Behavior types observed under each combination of (Z, D) in the binary case, after monotonicity has ruled out defiers.

Three behavior types and four observed expectations: the system is just identifiable enough to recover the LATE for compliers. Once we leave this tight 2×2 world, the number of behavior types grows quickly, and the identification arithmetic becomes more interesting.

6.3.2 Multiple Binary Instruments

Suppose we have two binary instruments z_1 and z_2, both valid (each independent of potential outcomes and inducing a first stage), and a binary treatment. The natural empirical move is to run 2SLS using both instruments. What does the resulting coefficient represent?

Angrist and Pischke (2009), Section 4.5.1, show that if monotonicity holds separately for each instrument, the 2SLS coefficient is a convex combination of the two instrument-specific LATEs:

Proposition 6.3.1 (2SLS with Multiple Binary Instruments, Section 4.5.1 in Angrist and Pischke, 2009)**.** *With two binary instruments z_1, z_2, both satisfying independence and monotonicity, the 2SLS estimand is*

$$\rho_{2SLS} = \psi \, LATE_1 + (1 - \psi) \, LATE_2, \qquad \psi \in (0, 1),$$

where $LATE_k$ is the LATE identified by instrument z_k alone. To express the weight, let the first stage fitted treatment be $\hat{D}_i = \pi_1 z_{1i} + \pi_2 z_{2i}$, where π_1 and π_2 are the first stage coefficients. Then, following Angrist and Pischke (2009, p. 131),

$$\psi = \frac{\pi_1 \operatorname{Cov}(D_i, z_{1i})}{\pi_1 \operatorname{Cov}(D_i, z_{1i}) + \pi_2 \operatorname{Cov}(D_i, z_{2i})} \in (0, 1),$$

so 2SLS gives more weight to the instrument with the stronger first stage.

The intuition is that 2SLS gives more weight to the instrument that does more work in the first stage. A strong instrument with a small LATE will pull the combined estimand toward that small LATE; a weak instrument with a large LATE will contribute less. This is reassuring and a little disquieting at the same time. Reassuring because 2SLS is at least a sensible weighted average; disquieting because the weights are determined by statistical first-stage strength, not by what we care about substantively. If LATE_1 and LATE_2 are similar in magnitude, the combination is informative. If they differ a

lot, ρ_{2SLS} can be hard to interpret as a single parameter.

There is also a deeper subtlety. With multiple instruments we have an overidentified system, and we can run the Hansen J-test for overidentifying restrictions (Chapter 5). The test asks whether the two instruments give statistically consistent estimates. Rejecting the test is often interpreted as evidence that at least one instrument is invalid. But with heterogeneous treatment effects, it could equally well mean that the two instruments are valid but identify different LATEs. The J-test cannot distinguish these two stories. In an MTE framework, the second story turns out to be a feature rather than a bug, because the two LATEs correspond to different windows of U_D.

6.3.3 Multivalued Treatment: The Average Causal Response

Now keep the instrument binary but let the treatment take ordered integer values $s \in \{0, 1, 2, \ldots, \bar{s}\}$. The canonical example is years of schooling. The instrument z might be a compulsory schooling law: $z = 1$ if the law applies, $z = 0$ otherwise. The treatment is the resulting years of schooling, which is naturally multivalued.

Angrist and Pischke (2009) state the relevant generalization, the Average Causal Response (ACR). The assumptions parallel the binary LATE assumptions:

Assumption 6.3.1 (ACR)**.** Assume that Y_{si} is the potential outcome of Y if treatment is s, and s_{zi} is the potential outcome of s if the instrument is z.

1. **Independence:** $\{Y_{0i}, Y_{1i}, \ldots, Y_{\bar{s},i}; s_{0i}, s_{1i}\} \perp z_i$. The instrument is as good as randomly assigned with respect to all potential outcomes and all potential treatments.
2. **First stage:** $E[s_{1i} - s_{0i}] \neq 0$. The instrument shifts the average level of treatment.
3. **Monotonicity:** $s_{1i} - s_{0i} \geq 0$ for all i (or weakly the reverse for all i). The instrument moves every agent's treatment in the same direction.

The monotonicity assumption now has bite: it requires us to have an ordered list of treatment values, so that the phrase "the instrument moves treatment in the same direction" is meaningful. For schooling this is natural; for unordered categorical treatments it is not.

Under these assumptions, the IV estimand has a clean interpretation.

Theorem 6.3.1 (Average Causal Response, Theorem 4.5.3 in Angrist and Pischke, 2009)**.** *Under independence, first stage, and monotonicity, the IV/Wald estimator with a binary instrument and ordered multivalued treatment satisfies*

$$\frac{E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0]}{E[s_i \mid z_i = 1] - E[s_i \mid z_i = 0]} = \sum_{s=1}^{\bar{s}} \omega_s \, E[Y_{s,i} - Y_{s-1,i} \mid s_{1i} \geq s > s_{0i}],$$

where the weights are

$$\omega_s = \frac{P[s_{1i} \geq s > s_{0i}]}{\sum_{j=1}^{\bar{s}} P[s_{1i} \geq j > s_{0i}]}.$$

Unpack this slowly. The quantity $Y_{s,i} - Y_{s-1,i}$ is the unit response, or stepwise treatment effect, of moving from treatment level $s-1$ to s for individual i. For each step, we average over the compliers whose treatment span crosses that step: agents for whom s_{0i} is below s and s_{1i} is at or above s. The weight ω_s is the share of all step-crossings that occur at step s. By construction the weights sum to one, so the estimand is a proper convex combination of stepwise effects.

This is the ACR: a weighted average of the unit causal responses, over the steps and over the compliers who cross each step. For schooling, an IV estimate using a compulsory schooling law identifies an average return per year of schooling, but only over the marginal years of schooling that the law actually shifts, and only for the compliers who shift on those margins. It is not the population average return per year, and it is not the return for any one fixed margin (say, the return to a four year college degree). What it is, and where it lies in the distribution of returns, depends on which margins the instrument moves.

If we have multiple binary instruments and multivalued treatment, the interpretation combines this case with the previous one: a weighted average across instruments, each of which contributes its own ACR. The weighting scheme becomes baroque quickly, and the practical advice is to be very explicit about what compliers and what margins your instrument is moving.

6.3.4 Multivalued Instrument and the Monotonicity Trap

Now consider the opposite case: binary treatment but multivalued instrument $z \in \{0, 1, 2, \ldots\}$. An example is the MTO experiment of Pinto (2015), where three voucher policies are randomly assigned. We will return to MTO in detail below. First, let us

see why this case is genuinely harder.

A natural first instinct is to decompose the multivalued instrument into binary dummies and run 2SLS. With $z \in \{0, 1, 2\}$ we could define

$$z_1 = \mathbf{1}[z = 1], \qquad z_2 = \mathbf{1}[z = 2],$$

and use both as instruments. Each dummy looks binary, so we are tempted to treat the resulting estimand as a weighted average of two LATEs, as in the multiple-IV case above. Unfortunately monotonicity can fail for the dummies even when it holds for z.

To see this, consider the group with $z_1 = 0$. These agents have either $z = 0$ or $z = 2$. Suppose the original instrument z is monotonic in the natural ordering, so that $D(z = 0) \leq D(z = 1) \leq D(z = 2)$. When z_1 flips from 0 to 1, agents starting from $z = 0$ move *up* to $z = 1$ (treatment weakly increases), but agents starting from $z = 2$ move *down* to $z = 1$ (treatment weakly decreases). The same change in z_1 pushes different agents in opposite directions in D. Monotonicity for z_1 is violated.

This is what we will call the *monotonicity trap*. It is not a small technical issue. It means that the IV estimate using these naive dummies is a weighted combination of effects that may have different signs, and the weights can even be negative. We cannot interpret the resulting estimand as any single average treatment effect for a clean subpopulation.

But is there a different dummy decomposition that could avoid the trap? The answer is yes: instead of indicator dummies, use cumulative dummies $\tilde{z}_1 = \mathbf{1}[z \geq 1]$ and $\tilde{z}_2 = \mathbf{1}[z \geq 2]$. With the natural ordering and monotonicity of z, each cumulative dummy is monotone in D, and the 2SLS estimand is a non-negative convex combination of the LATEs for the marginal step shifts. We leave the verification as a homework problem; it generalizes immediately to instruments with more than three values.

Even with cumulative dummies, however, monotonicity is no longer a free assumption. In the binary case it is essentially harmless: it says the instrument has the same sign for everyone. First, in the multivalued case, it must hold for each step of the instrument, which leads to a set of much stronger assumptions. Second, in many cases instrument values do not have a natural order, leading to difficulty in defining the steps. Third, even if we could decompose the IV successfully, the weighted average of these LATEs is hard to interpret, because for compliers responding to $z \geq 1$, some are also responding

to $z \geq 2$, which leads to overlap of the complier groups for each sub-LATE before taking the average. As a result, we have to step into the agent's optimization problem to illustrate what we are identifying here.

6.3.5 Revealed Preference and the Pinto MTO Example

6.3.5.1 Axioms of Revealed Preference

The natural framework for thinking about how budget shifts (from the instrument) change choices (the treatment) is revealed preference. Two axioms from microeconomics, the Weak Axiom of Revealed Preference (WARP) and the Strong Axiom of Revealed Preference (SARP), turn out to be exactly the tools we need.

Definition 6.3.1 (WARP, Definition 2.F.1 in Mas-Colell, Whinston, and Green, 1995)**.** The demand function $x(p, w)$ satisfies the Weak Axiom of Revealed Preference if, for any two price-wealth situations (p, w) and (p', w'):

$$p \cdot x(p', w') \leq w \text{ and } x(p', w') \neq x(p, w) \implies p' \cdot x(p, w) > w'.$$

In words: if some bundle is feasible at (p, w) but not chosen, and the agent chose a different bundle there, then that chosen bundle must not be affordable in the new situation. That is, $x(p, w)$ is revealed preferred to $x(p', w')$, written $x(p, w) \succsim_R x(p', w')$.

Definition 6.3.2 (SARP, Definition 3.J.1 in Mas-Colell, Whinston, and Green, 1995)**.** The demand function $x(p, w)$ satisfies the Strong Axiom of Revealed Preference if, for any sequence of price-wealth situations $(p^1, w^1), \ldots, (p^N, w^N)$ with $x(p^{n+1}, w^{n+1}) \neq x(p^n, w^n)$ for all $n \leq N-1$, we have $p^N \cdot x(p^1, w^1) > w^N$ whenever $p^n \cdot x(p^{n+1}, w^{n+1}) \leq w^n$ for all $n \leq N - 1$.

SARP is WARP plus transitivity. If x_1 is revealed preferred to x_2, which is revealed preferred to x_3, which is revealed preferred to $\ldots$, which is revealed preferred to x_N, then x_1 is revealed preferred to x_N. WARP rules out direct cycles between two bundles; SARP rules out cycles of any length.

For the IV identification problem, the SARP machinery lets us carry preferences across different policy environments (different instrument values). If we know an agent chose t over t' at one policy setting (with a corresponding budget), and the budget for t weakly expands and the budget for t' weakly shrinks under a new policy, then SARP

says the agent will still choose t over t' at the new budget set. This is exactly the kind of cross-policy restriction we need to rule out implausible behavior types and reduce the number of parameters we have to identify.

6.3.5.2 The MTO Setting

We illustrate this with the Moving to Opportunity experiment, analyzed by Pinto (2015). MTO randomized vouchers among low income families in five US cities, with the explicit goal of encouraging relocation to lower poverty neighborhoods. There are three random assignments (three values of the instrument):

- Control (z_1): no voucher.
- Experimental (z_2): voucher usable only for housing in a low poverty neighborhood.
- Section 8 (z_3): voucher usable for any housing.

And three possible relocation choices (three values of the treatment):

- $t = 1$: do not relocate.
- $t = 2$: relocate to a low poverty neighborhood.
- $t = 3$: relocate to a high poverty neighborhood.

The realized compliance pattern in MTO is shown in Figure 6.2, reproducing Figure 1 of Pinto (2015). Most families in the control group did not move; in the experimental group, voucher compliers all moved to low poverty neighborhoods (by construction); and in the Section 8 group, most voucher compliers moved to high poverty neighborhoods. Voucher non-compliers in the experimental and Section 8 groups behaved similarly to the control group, mostly not moving.

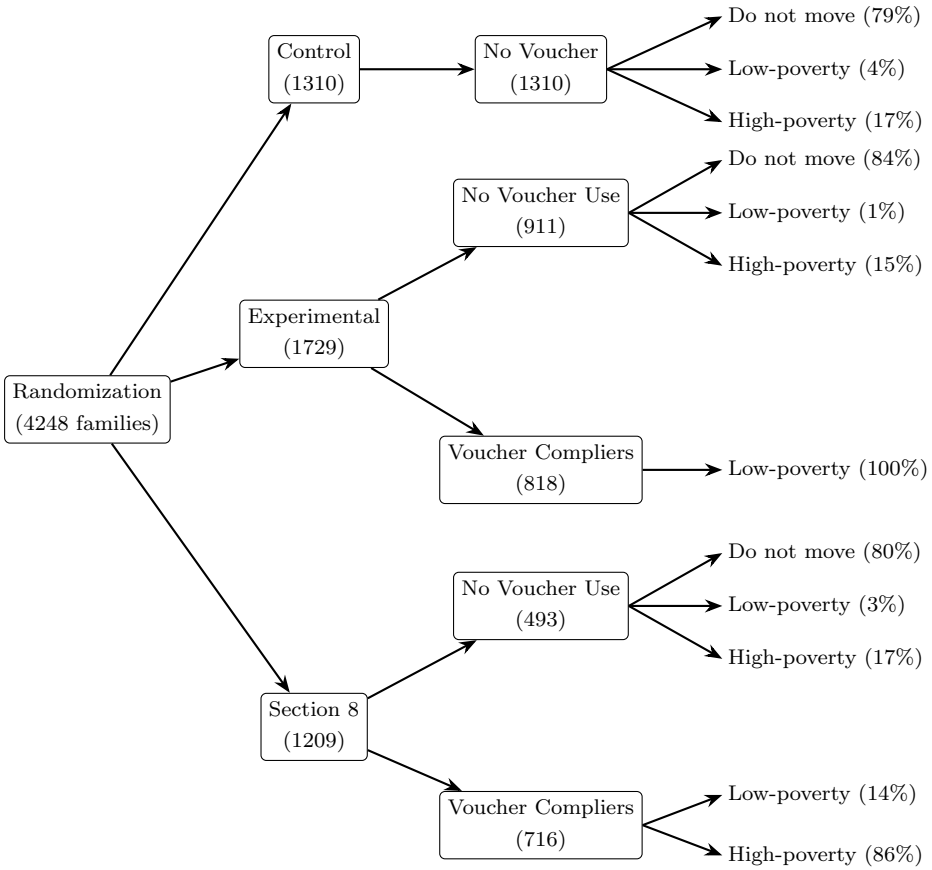


Figure 6.2: Neighborhood relocation by voucher assignment and compliance in MTO. Adapted from Figure 1 of Pinto (2015).

6.3.5.3 Why the Type Space Explodes

With three voucher values and three moving choices, an agent's behavior type is described by the triple

$$S_\omega = [C_\omega(z_1), C_\omega(z_2), C_\omega(z_3)],$$

the choice the family would make under each voucher. There are $3^3 = 27$ possible types. We observe nine conditional expectations $E[Y \mid z, t]$ for $z \in \{z_1, z_2, z_3\}$ and $t \in \{1, 2, 3\}$. A system of 9 equations with 27 behavior types cannot identify the underlying causal parameters without further structure. In general, with n choices and m instrument values, the number of behavior types is n^m and the number of observed expectations is $n \times m$. The behavior types grow exponentially in m; the observations grow linearly. There is no hope of identification without ruling out behavior types.

Luckily, the MTO design itself, together with rational choice, lets us rule out most

of them. Let $u_\omega(k,t)$ be the family's utility over consumption k and moving choice $t \in \{1,2,3\}$, and let $W_\omega(z,t)$ be the budget set the family faces if it chooses t under voucher z. The three policies translate into the following budget set restrictions.

Assumption 6.3.2 (A-1 and A-2 in Pinto, 2015)**.** For every family ω:

$$W_\omega(z_1,2) \subsetneq W_\omega(z_2,2) = W_\omega(z_3,2), \tag{6.1}$$

$$W_\omega(z_1,3) = W_\omega(z_2,3) \subsetneq W_\omega(z_3,3), \tag{6.2}$$

$$W_\omega(z_1,1) = W_\omega(z_1,2) = W_\omega(z_1,3) = W_\omega(z_2,1) = W_\omega(z_2,3) = W_\omega(z_3,1). \tag{6.3}$$

In words: (1) under the experimental or Section 8 vouchers, moving to a low poverty neighborhood expands the budget set, because the voucher subsidizes it. (2) Under the Section 8 voucher, moving to a high poverty neighborhood also expands the budget, but the experimental voucher does not subsidize this choice. (3) For any other combination (not moving, or moving to a place the voucher does not subsidize), the budget set is the same as under no voucher.

Combining these with SARP and the standard assumption of rational preferences, Pinto (2015) derives the following choice restrictions.

Lemma 6.3.1 (Lemma L-1 in Pinto, 2015)**.** *If preferences are rational, then under Assumptions A-1 and A-2:*

$$\begin{aligned}
&1.\ C_\omega(z_1) = 2 \Rightarrow C_\omega(z_2) = 2,\ C_\omega(z_3) \neq 1,\\
&2.\ C_\omega(z_1) = 3 \Rightarrow C_\omega(z_2) \neq 1,\ C_\omega(z_3) \neq 1,\\
&3.\ C_\omega(z_2) = 1 \Rightarrow C_\omega(z_1) = 1,\ C_\omega(z_3) \neq 2,\\
&4.\ C_\omega(z_2) = 3 \Rightarrow C_\omega(z_1) = 3,\ C_\omega(z_3) = 3,\\
&5.\ C_\omega(z_3) = 1 \Rightarrow C_\omega(z_1) = 1,\ C_\omega(z_2) = 1,\\
&6.\ C_\omega(z_3) = 2 \Rightarrow C_\omega(z_2) = 2.
\end{aligned}$$

The proof, which we omit, applies SARP repeatedly. The flavor is captured by case 1. If a family chooses to move to a low poverty neighborhood under the control voucher z_1, when no moving option is subsidized, then under the experimental voucher z_2 the budget for moving to a low poverty neighborhood weakly expands while the budgets for not moving and for moving to a high poverty neighborhood do not change. SARP

says the family should still prefer moving to a low poverty neighborhood, so $C_\omega(z_2) = 2$. Under z_3, the budgets for both kinds of moves expand, so the family will not choose not to move; but the family might now prefer moving to a high poverty neighborhood, so we cannot rule out $C_\omega(z_3) = 3$. We can only conclude $C_\omega(z_3) \neq 1$.

To go further, Pinto (2015) adds the standard assumption that neighborhoods are normal goods.

Assumption 6.3.3 (A-3 in Pinto, 2015)**.** For every family ω, and for $z, z' \in \{z_1, z_2, z_3\}$: if $C_\omega(z) = t$ and $W_\omega(z, t) \subsetneq W_\omega(z', t)$, then $C_\omega(z') = t$.

If the family chose t at a smaller budget set and the budget for t strictly expands under a different voucher, the family still chooses t. This rules out cases like $C_\omega(z_1) = 2, C_\omega(z_2) = 2, C_\omega(z_3) = 3$: the family chose moving to low poverty under both z_1 and z_2; relative to z_1 the budget for moving to low poverty has strictly expanded, so A-3 with $z = z_1$, $z' = z_3$ forces $C_\omega(z_3) = 2$.

Combining the rules from Lemma L-1 and Assumption A-3, the 27 a priori behavior types collapse to just 7 admissible types. With 9 observed expectations and 7 unknowns, the identification problem becomes manageable.

Lesson

We could not have collapsed 27 to 7 with statistical assumptions alone. Monotonicity in the naive "dummy decomposition" sense was not enough. What did the work was a sequence of economic assumptions: a budget set description of each policy, SARP applied to the underlying preferences, and the normal good restriction. When statistics runs out, microeconomics steps in. Do not throw away your knowledge from the first year Ph.D. microeconomics course.

6.4 Marginal Treatment Effect (MTE) Framework

6.4.1 Why MTE? External Validity

We have so far been generalizing LATE inside its original design-based framework. The other limitation we listed at the start of the chapter is external validity. LATE is the average treatment effect for compliers, where compliers are defined by the way agents respond to a specific instrument. If we change the policy, the complier subpopulation

changes too. Vietnam draft compliers and Vietnam volunteers are different groups; a hypothetical future draft lottery would identify yet another complier subpopulation.

The problem is that LATE is grouped by ex post behavior, not by ex ante characteristics. We would much rather have a treatment effect indexed by something that the agent brings to the policy, like ability, family background, or unobserved taste, because such ex ante categories are stable across policy environments. The MTE framework, developed by Heckman and Vytlacil (2005, 2007a,b), delivers exactly this. The trick is to write down an explicit model of how the agent decides to take treatment, and to use that model to translate ex post complier categories into ex ante categories indexed by an unobserved resistance to treatment.

6.4.2 The Latent Index Choice Model

We work with a binary treatment $D \in \{0, 1\}$ and a pair of potential outcomes

$$Y_1 = \mu_1(X, U_1), \tag{6.4}$$

$$Y_0 = \mu_0(X, U_0), \tag{6.5}$$

where X is a set of observed covariates and U_0, U_1 are unobserved factors that affect the outcomes in the two treatment states. The treatment choice is governed by a latent index model:

$$D^* = \mu_D(Z) - V, \qquad D = \mathbf{1}[D^* \geq 0]. \tag{6.6}$$

Z is a vector of instruments, including any elements of X, plus at least one variable that affects the choice but not the outcomes directly. V is the agent's unobserved resistance to treatment. The econometrician observes (X, Z, D, Y) but not (U_0, U_1, V). The three unobservables can be arbitrarily correlated with each other, so selection into treatment is allowed.

Assumption 6.4.1 (MTE Model)**.**

(A-1) $(U_0, U_1, V) \perp Z \mid X$. The instrument is independent of all unobservables conditional on X.

(A-2) $\mu_D(Z)$ is nondegenerate conditional on X. There is at least one element of Z that varies and is not contained in X, so the instrument actually moves the choice probability.

(A-3) The distribution of V is continuous.

(A-4) $E[|Y_1|]$ and $E[|Y_0|]$ are finite.

(A-5) $0 < \Pr(D = 1 \mid X) < 1$. Treatment status is not deterministically tied to X; both treated and untreated agents exist at every value of X.

(A-1) and (A-2) are the existence of a valid instrument, written in terms of the latent index. (A-3) gives us a continuous propensity score. (A-4) is a regularity condition for taking expectations. (A-5) is the overlap (common support) condition.

6.4.3 A Roy-Style Example

The latent index model is more concrete than it first appears. Consider the canonical Roy sorting model with two sectors: agriculture ($D = 0$) and a modern sector ($D = 1$). Let Y be a labor market payoff and let Z_1 be an observed relative cost of working in the modern sector. Assume additive separability in the outcome equations:

$$Y_1 = \mu_1(X) + U_1,$$
$$Y_0 = \mu_0(X) + U_0.$$

The agent's net utility from choosing the modern sector is the payoff differential minus the relative cost, minus an unobserved utility cost V_C:

$$D^* = \mu_1(X) + U_1 - [\mu_0(X) + U_0] - Z_1 - V_C, \qquad D = \mathbf{1}[D^* \geq 0].$$

Rearranging,

$$D^* = \underbrace{\mu_1(X) - \mu_0(X) - Z_1}_{\mu_D(Z)} + \underbrace{[U_1 - U_0 - V_C]}_{-V},$$

which matches the latent index form (6.6) with $V = -[U_1 - U_0 - V_C]$. Agents with high $\mu_1 - \mu_0$ (high observable return to the modern sector) or with high $U_1 - U_0$ (high unobservable return) sort into modern. This is positive sorting on returns. The instrument Z_1, an observed cost shifter, only enters $\mu_D(Z)$ and not the outcomes, so it satisfies the exclusion restriction.

6.4.4 Propensity Score and the Threshold U_D

The next step is to rewrite the choice equation in terms of the propensity score. Define the propensity score

$$P(Z) \equiv \Pr(D = 1 \mid Z, X) = \Pr(\mu_D(Z) > V \mid Z, X) = F_{V|X}(\mu_D(Z)),$$

where $F_{V|X}$ is the conditional CDF of V given X. The propensity score is just the probability that an agent with instruments Z takes treatment.

Apply the same CDF to the unobserved resistance:

$$U_D \equiv F_{V|X}(V) \sim \text{Uniform}[0, 1].$$

U_D is a uniform random variable that records the agent's quantile in the conditional distribution of V. It is also the threshold propensity score the agent has to pass to get treated when he or she draws resistance V. An agent with $U_D = 0.2$ has lower than average resistance to treatment; an agent with $U_D = 0.9$ has high resistance. Because $F_{V|X}$ is strictly monotonic (under continuity), V and U_D carry the same information, but U_D is normalized to lie in the unit interval.

Now compare the propensity score and the threshold:

$$\mu_D(Z) > V \iff \underbrace{F_{V|X}(\mu_D(Z))}_{P(Z)} > \underbrace{F_{V|X}(V)}_{U_D}.$$

The agent takes treatment if and only if the propensity score from Z exceeds the agent's quantile of resistance. We can therefore characterize each agent by the pair (X, U_D) rather than (X, V). The unobservable is now bounded in $[0, 1]$ and uniform by construction. This re-parameterization is the technical move that makes the rest of the framework work.

6.4.5 Vytlacil Equivalence

A central result is due to Vytlacil (2002): under assumptions (A-1) to (A-5), the additively separable latent index model is observationally equivalent to the LATE model of Imbens and Angrist (1994). In particular, the latent index model automatically implies monotonicity.

The intuition is that the choice equation $D^* = \mu_D(Z) - V$ is additively separable in Z and V. Given any two instrument values z and z',

$$D^*(z) - D^*(z') = \mu_D(z) - \mu_D(z'),$$

which has the same sign for every agent regardless of V. If $\mu_D(z) > \mu_D(z')$, every agent's latent utility for treatment goes up when we move from z' to z; nobody can go in the opposite direction. That is exactly monotonicity.

The equivalence works in both directions: the latent index model with (A-1) to (A-5) implies the LATE assumptions, and conversely the LATE assumptions can be represented by a latent index model. So MTE is not an alternative framework that replaces LATE; it is a richer parameterization of the same identification setup. The benefit is that the latent index gives us a continuous index U_D along which we can disaggregate the treatment effect.

6.4.6 Defining the Marginal Treatment Effect

Let $\Delta = Y_1 - Y_0$ denote the individual treatment effect. The conditional ATE is the population mean

$$\Delta^{\text{ATE}}(x) \equiv E[\Delta \mid X = x].$$

The Marginal Treatment Effect localizes this by conditioning on the unobservable as well.

Definition 6.4.1 (MTE, Heckman and Vytlacil, 2005)**.** The Marginal Treatment Effect at observable x and unobserved resistance u_D is

$$\Delta^{\text{MTE}}(x, u_D) \equiv E[Y_1 - Y_0 \mid X = x, U_D = u_D].$$

This is the average treatment effect for agents who share covariates x and quantile u_D of unobserved resistance. Equivalently, it is the average effect for agents who would be exactly indifferent between treatment and no treatment if they were assigned an instrument value z with $P(z) = u_D$. They are "marginal" in the textbook sense: at the threshold.

A few features make MTE attractive.

First, the MTE is indexed by ex ante characteristics. x describes the observable type and u_D describes the unobserved appetite for treatment. Neither depends on the particular instrument used in any one study. If we change the policy, the same agent has the same (x, u_D), and the same MTE value.

Second, the MTE traces out treatment effect heterogeneity along u_D. In a positive-selection Roy model, agents with low u_D are those most eager to be treated; they tend to have high gains. Agents with high u_D are reluctant to be treated and tend to have low or negative gains. The MTE curve as a function of u_D is typically decreasing in such settings, and a flat MTE corresponds to homogeneous treatment effects.

Third, and most importantly for empirical work, the MTE is a deep structural object that does not depend on the instrument. LATE, ATE, ATT, ATUT, and even the OLS estimand can all be written as weighted averages of $\Delta^{\text{MTE}}(x, u_D)$ with different weights. If we know the MTE, we know all of these.

6.4.7 MTE as a Unified Framework

The general representation is

$$\text{TE}(j) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_j(x, u_D)\, du_D,$$

where j indexes the treatment parameter (ATE, ATT, LATE, etc.) and ω_j is the parameter-specific weight. Four leading cases:

ATE. $\omega_{\text{ATE}}(x, u_D) = 1$. The ATE simply integrates the MTE over the full range of U_D.

ATT (Treatment on the Treated). The weight ω_{ATT} puts more mass on agents with low u_D, who are over-represented among the treated under positive selection. The ATT is typically larger than the ATE in a positive sorting model.

ATUT (Treatment on the Untreated). The weight ω_{ATUT} puts more mass on agents with high u_D, who are over-represented among the untreated. Under positive selection, ATUT is below the ATE.

LATE. For an instrument that shifts the propensity score from u'_D to $u_D > u'_D$, the

LATE is the average of the MTE over the induced window:

$$\begin{aligned}\text{LATE} &= E[Y_1 - Y_0 \mid X = x, D(z) = 1, D(z') = 0] \\ &= E[Y_1 - Y_0 \mid X = x, u'_D < U_D \leq u_D] \\ &= \frac{1}{u_D - u'_D} \int_{u'_D}^{u_D} \Delta^{\text{MTE}}(x, u)\, du. \end{aligned} \tag{6.7}$$

A LATE is the average of the MTE over the window of U_D that the instrument actually moves. Different instruments move different windows, which is exactly why different instruments give different LATEs.

Figure 6.3 shows a stylized picture, recreating Figure 1 of Heckman and Vytlacil (2005), of the MTE together with the ATE, the ATT weight, and the ATUT weight as functions of u_D in a positive-selection Roy model. The MTE declines in u_D (high resistance agents have low gains), the ATT weight is concentrated at low u_D, the ATUT weight is concentrated at high u_D, and the ATE is the unweighted area under the MTE curve.

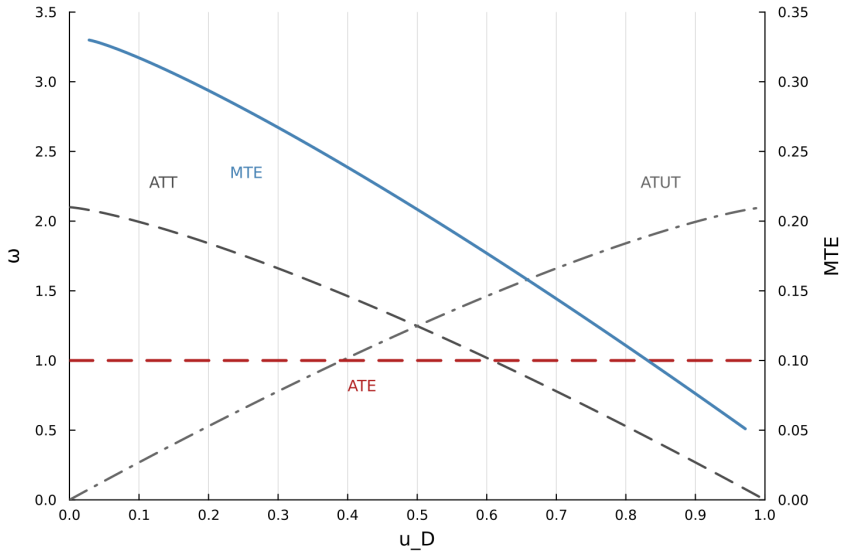


Figure 6.3: Stylized MTE curve with ATE reference line and ATT/ATUT weights in a positive-selection Roy model. Recreated from Figure 1 of Heckman and Vytlacil (2005).

The picture makes clear that the various treatment parameters are not competitors but views of the same object. They differ only in which part of the MTE curve they emphasize, which is determined by the weight function. The MTE itself is policy

invariant: the underlying schedule does not move as the instrument changes.

6.4.8 Estimation via the Local Instrumental Variable

How do we estimate the MTE from data? The key insight, again from Heckman and Vytlacil (2005), is that the MTE can be recovered as the derivative of a conditional mean function. Suppress the conditioning on X for clarity and write the propensity score as $P(Z)$. Define the **Local Instrumental Variable (LIV)** as

$$\Delta^{\text{LIV}}(p) \equiv \frac{\partial E[Y \mid P(Z) = p]}{\partial p}.$$

LIV is the local response of the outcome to a small change in the propensity score.

Theorem 6.4.1 (MTE-LIV Equivalence, Heckman and Vytlacil, 2005)**.** *Under assumptions (A-1) to (A-5),*

$$\Delta^{MTE}(p) = \Delta^{LIV}(p) = \frac{\partial E[Y \mid P(Z) = p]}{\partial p}.$$

The intuition is clean. A small policy nudge that raises the propensity score from p to $p + dp$ only changes the treatment status of marginal agents whose threshold is exactly $u_D = p$. Among those marginal agents, the change in Y is precisely their treatment effect, which is $\Delta^{\text{MTE}}(p)$. Aggregating, the derivative of $E[Y \mid P]$ at p delivers the MTE at p.

In practice the estimation runs in three steps.

Step 1. Estimate a treatment choice model. A probit or logit regression of D on Z delivers $\hat{P}(Z)$, the estimated propensity score.

Step 2. Estimate $E[Y \mid X, P(Z)]$ as a function of $\hat{P}$. This can be done flexibly: a series expansion in $\hat{P}$, a kernel smoother, a local polynomial. With a linear specification one obtains a constant MTE; with a flexible specification one obtains a nontrivial MTE curve.

Step 3. Take the derivative with respect to $\hat{P}$. If the second stage uses a polynomial of order K in $\hat{P}$, the MTE is given by the derivative polynomial. With non-parametric smoothing, the derivative is read off the slope of the conditional mean function.

Alternatively, the entire model can be estimated parametrically. Kline and Walters

(2016) write a fully parametric MTE model and estimate the joint distribution of (U_0, U_1, V) via maximum likelihood. The parametric approach is more efficient when the assumptions hold and is the natural choice when the goal is policy simulation.

6.4.9 Implementation: The `mtefe` Stata Package

The MTE framework can be implemented in Stata using the user written package `mtefe` by Martin Andresen. `mtefe` fits a parametric or semi-parametric MTE model and returns a battery of treatment effect estimates: ATE, ATT, ATUT, LATE, and Policy Relevant Treatment Effects (PRTE) for hypothetical policies. The package also produces tests for observable and unobserved heterogeneity. Typical output for a wage equation with experience controls and district fixed effects, with a district-level cost shifter as the instrument, includes effect estimates of the form

- ATE around 0.33,
- ATT around 0.54 (treated agents have above average gains, consistent with positive sorting),
- ATUT around 0.12 (untreated agents have lower gains),
- LATE close to the ATE in this stylized example.

Figure 6.4 illustrates the typical `mtefe` output. The left panel shows the common support of the propensity score for treated and untreated subsamples; the right panel shows the estimated MTE as a function of u_D together with a 95% confidence band and the ATE reference line.

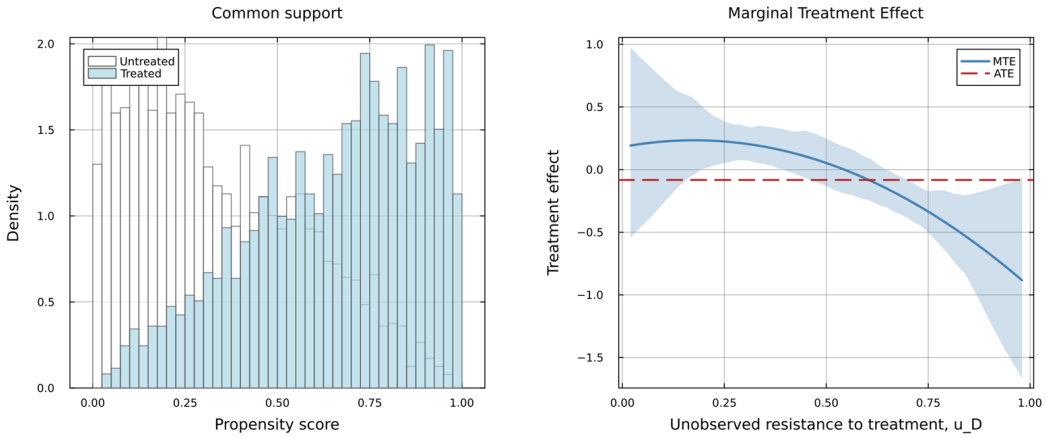


Figure 6.4: Stylized `mtefe` output. Left: propensity score density by treatment status. Right: estimated MTE curve with 95% confidence band and ATE reference line. The downward sloping MTE indicates positive selection on unobserved gains.

6.5 Application: Kline and Walters (2016)

Kline and Walters (2016) apply the MTE framework to evaluate the Head Start program, an early childhood education program for low income families in the United States. Their study is an instructive illustration of how the MTE delivers what neither a single LATE nor an observational comparison can deliver on its own.

Head Start enrollment looked very different in two kinds of studies. Observational comparisons of Head Start participants to non-participants showed large positive effects on later outcomes. The Head Start Impact Study (HSIS), a randomized control trial, showed small and often statistically insignificant effects. The natural inference is that observational studies suffered from selection bias and the RCT is closer to the truth. Kline and Walters (2016) argue this inference misses the point.

The reason is that the treatment in Head Start is not really binary. Families have three options: no preschool program at all, Head Start, or a close substitute (typically another subsidized or private preschool). Observational comparisons effectively contrast Head Start with the "no program" option. The RCT contrasts the assigned Head Start option with whatever counterfactual the control group can find, which often includes another preschool program. When close substitutes are available, the LATE identified by the RCT is the effect of Head Start versus the average alternative, not the effect of Head Start versus nothing. The two studies estimate different parameters and the

apparent contradiction dissolves.

Kline and Walters (2016) formalize this with a three valued treatment (no program, other program, Head Start) and use the MTO machinery from earlier in this chapter to characterize the admissible behavior types under the HSIS lottery. They estimate a parametric MTE model and report a rich set of causal parameters: ITT, LATE, and several MTE based objects that are externally valid in the sense that their interpretation does not depend on the specific compliance pattern of the HSIS sample.

The headline policy result is that, once we account for substitution from other preschool programs to Head Start, the program is far more cost effective than the small RCT impact would suggest. From the government's perspective, a family who switches from a different subsidized program to Head Start carries with them a savings on the alternative program. This savings, combined with the (modest) marginal benefit of Head Start over the substitute, gives Head Start a positive net social return.

The methodological lesson is the one we have been building toward. A single LATE estimate, no matter how cleanly identified, can be misleading when the treatment effect is heterogeneous and policy environments change the composition of compliers. The MTE framework lets us write down the policy question in terms of stable structural parameters, estimate those parameters from one experiment, and use them to forecast the effect of related but distinct policies.

6.6 Conclusion

LATE is internally valid and elegantly tied to a research design, but it leaves two questions open. What does 2SLS identify when the instrument or the treatment is not binary? And what does a LATE estimated under one policy tell us about a different policy?

The first half of this chapter addressed the first question by working through three cases: multiple binary instruments (a ψ-weighted average of instrument-specific LATEs), multivalued ordered treatment with a binary instrument (the Average Causal Response of stepwise effects), and multivalued instrument (where naive dummy decomposition breaks monotonicity). The multivalued instrument case forced us to import economic theory through WARP and SARP, and the MTO example of Pinto (2015) showed concretely how budget set restrictions plus rational choice cut the behavior type space

down from $3^3 = 27$ to a tractable 7.

The second half introduced the MTE framework of Heckman and Vytlacil (2005, 2007a,b). By writing down an explicit latent index choice model and re-parameterizing the unobserved resistance to lie in the unit interval, we obtain a treatment effect $\Delta^{\text{MTE}}(x, u_D)$ that is indexed by ex ante observables and ex ante unobservables, and that does not depend on which instrument we use. ATE, ATT, ATUT, LATE, IV, and OLS estimands are all weighted averages of the MTE. We can estimate the MTE through the Local Instrumental Variable, which is the derivative of $E[Y \mid P(Z)]$ with respect to the propensity score, using either non-parametric smoothing or fully parametric maximum likelihood. Kline and Walters (2016) illustrate the framework with Head Start.

The thread running through both halves is the same: treatment selection is intrinsic to any IV exercise, and ignoring it means leaving information on the table. Importing economic theory back into IV regularizes identification in hard cases, generalizes interpretation to richer policy settings, and makes the resulting estimates more useful for forecasting future policies.

Homework Problems

1. **Cumulative-Dummy Decomposition of a Multivalued Instrument.**

   Suppose the instrument z takes values in $\{0, 1, 2\}$ and the treatment D is binary. Assume the underlying instrument is monotonic, so $D(z = 0) \leq D(z = 1) \leq D(z = 2)$ for every agent.

   (a) Define the indicator dummies $z_1 = \mathbf{1}[z = 1]$ and $z_2 = \mathbf{1}[z = 2]$. Show that they do not jointly satisfy monotonicity for D, and identify a behavior pattern that violates monotonicity for z_1.

   (b) Now define cumulative dummies $\tilde{z}_1 = \mathbf{1}[z \geq 1]$ and $\tilde{z}_2 = \mathbf{1}[z \geq 2]$. Show that each $\tilde{z}_k$ is monotonic in D.

   (c) Argue that 2SLS using these cumulative dummies as instruments identifies a non-negative weighted average of the two stepwise LATEs: the LATE for the $z = 0 \to 1$ shift and the LATE for the $z = 1 \to 2$ shift.

   (d) Generalize: describe the cumulative dummy decomposition for an instru-

ment with $K+1$ values.

2. **Choice Restrictions in the Pinto MTO Example.**

   Verify two of the six implications in Lemma L-1 of Pinto (2015), restated in the notes. You may pick any two of the six. For each, state the relevant budget set comparisons under Assumptions A-1 and A-2, and explain how SARP delivers the implication.

3. **LATE as the Integral of the MTE.**

   Consider the latent index model in Section 6.4 with all of (A-1) to (A-5) satisfied. Let z and z' be two instrument values with $P(z) > P(z')$.

   (a) Show that the LATE identified by the instrument moving from z' to z equals

   $$\text{LATE} = \int_{P(z')}^{P(z)} \Delta^{\text{MTE}}(u)\, du \Big/ [P(z) - P(z')].$$

   (Note the normalization by the width of the propensity-score window.)

   (b) Suppose the MTE is constant at $\Delta^{\text{MTE}}(u) = c$ for all u. Show that every LATE equals c, and that the ATE, ATT, and ATUT also equal c.

Practice Example: Weighted LATE with Two Binary Instruments

This practice example illustrates the multiple binary instruments case from Section 6.3.2. We simulate a population in which two binary instruments identify two distinct LATEs, and we verify that the 2SLS coefficient is a convex combination of them, weighted by the relative first stage strengths.

Data generating process. Consider $N = 5000$ agents, each with a binary treatment D and a continuous outcome Y. Two binary instruments z_1 and z_2 are independently assigned with $\Pr(z_k = 1) = 0.5$. Each instrument shifts a distinct subgroup of agents from $D = 0$ to $D = 1$. Subgroup A consists of 30% of agents who respond only to z_1 with true treatment effect $\beta_A = 1.0$. Subgroup B consists of 30% of agents who respond only to z_2 with true treatment effect $\beta_B = 3.0$. The remaining 40% are never-takers or always-takers and do not respond to either instrument. The instruments z_1 and z_2 are

independent of potential outcomes by construction.

Predicted behavior. The Wald estimator using z_1 alone should recover $\text{LATE}_1 = \beta_A = 1.0$; using z_2 alone, $\text{LATE}_2 = \beta_B = 3.0$. The 2SLS estimator using both should produce

$$\rho_{\text{2SLS}} = \psi \cdot 1.0 + (1 - \psi) \cdot 3.0,$$

with ψ determined by the first stage strengths. Because the two subgroups are the same size, ψ should be close to 0.5 and the 2SLS estimate close to 2.0.

Analysis. The Stata do-file (`src/chapter 6/ch6_iv_beyond_late.do`) performs the following steps:

1. Simulate the population with two distinct complier subgroups.
2. Compute the two Wald estimators using each instrument separately. Confirm $\text{LATE}_1 \approx 1$ and $\text{LATE}_2 \approx 3$.
3. Run 2SLS using both instruments jointly. Verify the coefficient is a convex combination of the two LATEs. Also compute Hansen's J statistic for the overidentifying restriction from a two step GMM fit, to see that the test can reject even though both instruments are valid.
4. Recover the implied weight $\hat{\psi}$ by inverting the convex combination formula, and report it alongside the first stage F statistics.
5. Discuss interpretation: 2SLS gives more weight to the instrument with the stronger first stage.

Key lessons.

- With heterogeneous treatment effects and multiple instruments, 2SLS does not identify any one LATE: it identifies a convex combination.
- The combination weights are determined by first stage strength, not by what the researcher cares about. This is reassuring and slightly disquieting at the same time.
- In an MTE framework, the two LATEs correspond to different windows of U_D. The fact that they differ is not evidence of an invalid instrument; it is the heterogeneous structure of treatment effects.

The do file is located at `src/chapter 6/ch6_iv_beyond_late.do` and the simulated

dataset at `data/chapter 6/ch6_iv_beyond_late.dta`.

Chapter 7

Bartik Instruments

7.1 Introduction

In the previous two chapters we developed the general theory of instrumental variables and its extensions beyond the standard LATE framework. In this chapter we zoom in on one particular instrument that is widely used in applied economics: the *Bartik instrument*, also known as the *shift-share instrument*, or SSIV for short. The idea is old but its formal econometric treatment is recent.

The construction is deceptively simple. To instrument a location level endogenous variable such as local employment growth or a local migration inflow, we take a weighted sum of national industry level (or origin country level) shocks, where the weights are baseline shares of each industry or origin group in that location. The intuition is that a location that specialized in a particular industry at some earlier date should be disproportionately affected by a subsequent national shock to that industry, and this differential exposure gives us variation that is arguably exogenous to unobserved local factors.

Bartik instruments show up in a wide range of settings. In labor economics they are used to instrument for local labor demand or for the local inflow of immigrants (Card, 2009). In trade they are used to measure the exposure of US commuting zones to Chinese import growth (Autor et al., 2013). In urban and regional economics they are the workhorse for any question that requires a source of exogenous variation in local shocks.

Two natural questions arise. *How should we use this instrument correctly?* And *what exactly is the identification assumption?* The answer turns out to be more subtle than one might expect from the simplicity of the construction. There are actually two coherent frameworks that give a shift-share instrument a rigorous causal interpretation, and they are not equivalent in general. Goldsmith-Pinkham et al. (2020), henceforth GPSS, treat the shares as the identifying instrument and the shifts as weights. Borusyak et al. (2022), henceforth BHJ, do the opposite: the shifts are the identifying instrument and the shares are weights. Which framework applies to a given application depends on where the exogeneity actually comes from, and reasonable people can and do disagree about that in any particular setting. The good news is that a valid Bartik design can be defended by verifying the assumptions of either framework.

This chapter walks through both frameworks in turn. Section 7.2 sets up two motivating examples that will accompany us through the chapter. Section 7.3 develops the GPSS view, including the Bartik-GMM equivalence, the exposure design interpretation, the Rotemberg decomposition into single-industry instruments, and the empirical suggestions that follow. Section 7.4 develops the BHJ view, including the shock level equivalence, the quasi-random shock assumption, and the inference concerns that come with a shift-share error structure. Section 7.5 compares the two frameworks and gives guidance on which to use in different settings. Section 7.6 illustrates with the Autor et al. (2013) China shock application, which both papers use as their running example. Section 7.7 concludes.

7.2 Motivating Examples

7.2.1 Card (2009): Immigration and Inequality

Let us start with two motivating examples of the Bartik instrument. Card (2009) studies the effect of the immigrant share of the local labor force on the native-immigrant wage gap. The estimating equation is:

$$y_l = \beta_0 + \beta_1 \ln x_l + \beta_2 C_l + \epsilon_l, \tag{7.1}$$

where l indexes location (metropolitan area), y_l is the log wage gap between immigrants and natives, x_l is the ratio of immigrant labor to native labor, and C_l collects location level controls.

The immigrant ratio x_l is endogenous. Any positive local productivity shock is likely to raise both the wage gap (through skill sorting, occupational upgrading, and so on) and the immigrant inflow (immigrants respond to labor demand). OLS is therefore biased. Card constructs a shift-share instrument using data for 1980, 1990, and 2000:

$$B_l = \sum_k Z_{lk,1980} \cdot g_k, \tag{7.2}$$

$$Z_{lk,1980} = \left(N_{lk,1980}/N_{k,1980}\right) \times \left(1/P_{l,2000}\right), \tag{7.3}$$

where k indexes home country, $N_{lk,1980}$ is the number of immigrants from country k living in location l in 1980, $N_{k,1980}$ is the total US population of immigrants from k in 1980, $P_{l,2000}$ is the population of l in 2000, and g_k is the total inflow of immigrants from country k into the United States between 1990 and 2000. The share $Z_{lk,1980}$ measures the base year exposure of location l to origin country k; the shift g_k measures the size of the national wave from k.

The identification argument runs on two legs. *Relevance* holds because immigrants tend to cluster by origin. Locations that hosted a large Chinese community in 1980 (San Francisco, say) will receive a disproportionate share of the 1990 to 2000 Chinese inflow because new arrivals move to established communities. *Exclusion* holds if the location specific exposure to national origin country flows affects local wages only through the resulting change in the immigrant share and not through other local channels. This second leg is the substantive assumption and it deserves scrutiny.

The construction of B_l decomposes the local immigrant inflow into a location-origin country combination, weighted by baseline local shares and multiplied by national origin country flows. This is a *local share* times *national shift* object, hence the name shift-share, or Bartik, instrument.

7.2.2 Autor, Dorn, and Hanson (2013): The China Shock

The other running example we use in this chapter is Autor et al. (2013), whose China shock paper is one of the most cited applications of a shift-share instrument in the past two decades. The question is: what is the effect of the surge in Chinese imports on local labor markets in the United States? They construct a location level import

exposure variable

$$\Delta \text{IPW}_{it} = \sum_j \frac{L_{ijt}}{L_{jt} \cdot L_{it}} \Delta M_{jt},$$

where i indexes commuting zone, j indexes manufacturing industry, and t indexes year. Here L_{ijt} is employment in region i and industry j, L_{jt} is total US employment in industry j, L_{it} is total employment in region i, and ΔM_{jt} is the growth in US imports from China in industry j.

The structure is again a share times a shift. The local industry employment share L_{ijt}/L_{it} is the local exposure to industry j; the national import growth per worker $\Delta M_{jt}/L_{jt}$ is the industry shift. Locations that specialized in industries that were subsequently hit hard by Chinese import competition should have experienced larger declines in local labor market outcomes, and this differential exposure is what the paper exploits.

We will come back to the China shock in Section 7.6, after we have the two frameworks in hand.

7.3 Goldsmith-Pinkham, Sorkin, and Swift (2020): Share as Instrument

7.3.1 Setup and Two Identities

We now step away from the specific examples and set up a general framework. Following Goldsmith-Pinkham et al. (2020), consider the following regression:

$$y_{lt} = D'_{lt}\rho + x_{lt}\beta_0 + \epsilon_{lt}, \tag{7.4}$$

where l indexes location, t indexes time, D_{lt} is a vector of controls (which may include location and time fixed effects), x_{lt} is an endogenous variable, and β_0 is the parameter of interest. In many applications x and y are growth rates; when they are levels, the researcher should control for location and time fixed effects so that identification works off within-location changes rather than cross sectional levels.

Usually, x_{lt} is correlated with ϵ_{lt}, so we need an instrument. The Bartik construction comes from two algebraic identities. First, we can decompose the location-time level

variable x_{lt} into a sum over some finer categorization (industries in the Autor-Dorn-Hanson example, origin countries in the Card example) of a share times a location-category level growth:

Identity 1: Share-Growth Decomposition

$$x_{lt} = Z_{lt} G_{lt} = \sum_{k=1}^{K} z_{lkt}\, g_{lkt},$$

where z_{lkt} is the share of category k in location l at time t, and g_{lkt} is the location-category growth rate at t.

This is not an assumption. It is just how one writes a sum. The share z_{lkt} sums to one across k, and g_{lkt} is defined so that the identity holds.

Second, we can decompose the location-category growth into a national industry component plus a residual local component:

Identity 2: National-Local Decomposition

$$g_{lkt} = g_{kt} + \tilde{g}_{lkt},$$

where g_{kt} is the national growth rate of category k at t and $\tilde{g}_{lkt}$ is the location-category deviation from the national trend.

Combining the two identities, one gets

$$x_{lt} = \sum_{k} z_{lkt}\big(g_{kt} + \tilde{g}_{lkt}\big) = \sum_{k} z_{lkt} g_{kt} + \sum_{k} z_{lkt} \tilde{g}_{lkt}.$$

Now here comes the substantive move: fix the shares at a baseline period 0 (drop the time subscript on shares), and drop the local shock component $\tilde{g}_{lkt}$ from the sum. What is left is the Bartik instrument:

Definition 7.3.1 (Bartik Instrument, Goldsmith-Pinkham et al., 2020)**.**

$$B_{lt} = Z_{l0} G_t = \sum_{k} \underbrace{z_{lk0}}_{\text{Share}} \underbrace{g_{kt}}_{\text{Shift}}.$$

The first factor is the baseline share of category k in location l; the second is the national growth of category k at t.

The virtue of this construction is that both factors are, at least on the surface, more plausibly exogenous than x_{lt} itself. The baseline share is fixed in the past and cannot respond to future shocks. The national shift is an aggregate object that averages out any single location's contribution. In some cases, we can also use leave-one-out growth rate to exclude location l when calculating the national growth. So it is close to exogenous with respect to any one location's unobservables. Whether this intuition survives careful scrutiny is exactly what the identification analysis has to establish.

7.3.2 Two Special Cases

Before we state the general result, it helps to work through two tractable special cases that already convey the main intuition.

7.3.2.1 Case 1: Two Industries, One Period

Suppose we have only two categories ($K = 2$) and one time period. Since shares sum to one, $z_{l2} = 1 - z_{l1}$, so

$$B_l = z_{l1}g_1 + z_{l2}g_2 = g_2 + (g_1 - g_2)z_{l1}.$$

Now consider the first stage regression of x_l on B_l:

$$x_l = \gamma_0 + \gamma B_l + \eta_l = \underbrace{(\gamma_0 + \gamma g_2)}_{\text{constant}} + \underbrace{\gamma(g_1 - g_2)}_{\text{coefficient}} z_{l1} + \eta_l.$$

Because g_1 and g_2 are constants across locations, using the Bartik instrument B_l in 2SLS is algebraically identical to using the single share z_{l1} as the instrument. All the identifying variation is coming from the cross-location variation in z_{l1}; the shifts g_1 and g_2 just rescale the coefficient. This is important: in this simplest case, the identification comes from the share.

7.3.2.2 Case 2: Two Industries, Two Periods

The next case adds time. With $K = 2$ and $T = 2$,

$$B_{lt} = g_{1t}z_{l10} + g_{2t}z_{l20} = g_{2t} + (g_{1t} - g_{2t})z_{l10}.$$

Assume we control for time fixed effects. The first stage becomes

$$x_{lt} = \tau_t + \gamma B_{lt} + \eta_{lt} = \underbrace{(\tau_t + g_{2t}\gamma)}_{\tilde{\tau}_t} + z_{l10}(g_{1t} - g_{2t})\gamma + \eta_{lt}.$$

Now the coefficient $g_{1t} - g_{2t}$ is not constant: it varies with t. So the single-share reduction does not go through as cleanly. Using the indicator function $\mathbf{1}(\cdot)$, we can write

$$g_{1t} - g_{2t} = \mathbf{1}(t = 1)(g_{11} - g_{21}) + \mathbf{1}(t = 2)(g_{12} - g_{22}),$$

and the first stage becomes

$$x_{lt} = \tilde{\tau}_t + z_{l10}\mathbf{1}(t = 1)\underbrace{(g_{11} - g_{21})\gamma}_{\tilde{\gamma}_1} + z_{l10}\mathbf{1}(t = 2)\underbrace{(g_{12} - g_{22})\gamma}_{\tilde{\gamma}_2} + \eta_{lt}.$$

The regression now has time fixed effects and two interactions between the baseline share z_{l10} and each time dummy. The identification exploits both the cross-section variation in z_{l10} and the time variation in the industry growth differential $g_{1t} - g_{2t}$.

Here, the second special case has a **policy exposure research design reading**. The growth rate difference $g_{1t} - g_{2t}$ plays the role of a policy shock size at time t: it says how differently industry 1 grew relative to industry 2 in period t. The baseline share z_{l10} plays the role of exposure to that shock: locations with a high share of industry 1 are more exposed. The question the regression asks is whether locations with higher exposure experience different changes in x following the shock.

In the limit case of a pre-shock period ($t = 1$) with no differential growth ($g_{11} - g_{21} = 0$) and a post-shock period ($t = 2$) with a positive differential growth, this is a DID specification: $\tilde{\gamma}_1 = 0$ is a parallel-trend condition, and $\tilde{\gamma}_2 > 0$ measures the effect of the shock.

So Bartik with a baseline share is really a diff-in-diff on differential exposure. Locations differ in their exposure to a common industry shock, and we compare how outcomes evolve for high-exposure and low-exposure locations. That is the identifying variation.

7.3.3 Bartik-GMM Equivalence

Now let us state the general result. Assume K industries and one period (the $T > 1$ case is a straightforward extension). Stack observations into matrix form: Y and X

are $L \times 1$, Z is $L \times K$ (with entry z_{lk0} in row l column k), and G is $K \times 1$ (with g_k in row k). Let $M_D = I - D(D'D)^{-1}D'$ be the annihilator matrix that partials out D, and write $X^{\perp} = M_D X$, $Y^{\perp} = M_D Y$, $B^{\perp} = M_D B$ where $B = ZG$ is the $L \times 1$ Bartik instrument.

Tips: The annihilator matrix M_D is used to partial out the impact of control variable D. For any variable matrix X, $M_D X$ gives the OLS regression residual of X on D. Then by Frisch–Waugh–Lovell Theorem, we can estimate the IV/OLS coefficient of interest by running the regressions using residualized X and Y. For more details, please refer to Hansen (2022).

Proposition 7.3.1 (Bartik-GMM Equivalence, Proposition 1 in Goldsmith-Pinkham et al., 2020)**.** *Define the Bartik and GMM estimators*

$$\hat{\beta}_{Bartik} = \frac{B'Y^{\perp}}{B'X^{\perp}}, \qquad \hat{\beta}_{GMM} = \frac{X^{\perp\prime}ZWZ'Y^{\perp}}{X^{\perp\prime}ZWZ'X^{\perp}}.$$

If the GMM weight matrix is $W = GG'$, then $\hat{\beta}_{Bartik} = \hat{\beta}_{GMM}$.

The interpretation is clean. The Bartik instrument $B = ZG$ is a linear combination of the K share instruments $Z_1, \ldots, Z_K$ with weights given by the shifts $g_1, \ldots, g_K$. Running 2SLS with the single combined instrument B is exactly the same as running an overidentified GMM with the K separate share instruments $Z_1, \ldots, Z_K$, using a specific weight matrix $W = GG'$. **In the GPSS framework, therefore, the identifying instruments are the shares, and the shifts serve as the weights that combine them.**

7.3.4 Consistency and What Exogeneity Really Requires

For the Bartik estimator to be consistent, we need the standard IV conditions to hold at the level of the share instruments. Goldsmith-Pinkham et al. (2020) state the following.

Assumption 7.3.1 (Relevance)**.** The instrument is correlated with the endogenous regressor after partialling out controls.

Assumption 7.3.2 (Strict Exogeneity of Shares)**.** For every category k with $g_k \neq 0$,

$$E[\epsilon_{lt} z_{lk0} \mid D_{lt}] = 0.$$

Proposition 7.3.2 (Consistency, Proposition 2 in Goldsmith-Pinkham et al., 2020)**.** *Under Assumptions 7.3.1 and 7.3.2, and asymptotics in the location dimension ($L \to \infty$ with T, K fixed),*

$$\text{plim}\, \hat{\beta}_{Bartik} = \beta_0.$$

The key assumption is exogeneity of the shares. When is it plausible? Suppose y_{lt} is a level rather than a growth rate. Then ϵ_{lt} absorbs all unobserved determinants of the level, including any persistent unobserved characteristics of location l. The baseline share z_{lk0} is almost certainly correlated with these persistent local characteristics. San Francisco's 1980 share of Chinese immigrants is not independent of San Francisco's unobserved 2000 amenities, culture, or productivity. In levels, Assumption 7.3.2 fails.

Levels vs Changes: A Practical Warning

Initial shares are typically *not* mean independent of unobserved outcome *levels*. Persistent local characteristics that shape 1980 shares also shape 2000 outcomes.
Initial shares are more plausibly mean independent of unobserved outcome *changes*. Once we difference out the level, we are left with idiosyncratic shocks over the sample period that are less likely to be tied to the baseline industry mix.
When using a Bartik instrument, either use a growth variable as the outcome, or control for location and time fixed effects. Never run a Bartik IV in cross-sectional levels without location fixed effects.

Substantively, Assumption 7.3.2 is an *exposure design* assumption, close in spirit to parallel trends in DID. Locations that were more exposed to an industry are more affected by shocks to that industry, but the assumption says there is no systematic difference in *other* unobserved shocks between high-exposure and low-exposure locations. Think of Shanghai and Hong Kong versus Shenyang and Wuhan. Shanghai and Hong Kong specialize in finance, so a global financial crisis hits them harder. For the Bartik instrument to be valid, we need to assume that in the absence of the crisis shock, the outcome trends in the four cities would have been parallel. Any unobserved shocks that hit finance-heavy cities differentially would invalidate the design.

7.3.5 The Rotemberg Decomposition

The GMM equivalence result tells us that a Bartik IV is really a combination of many single-industry instruments. A natural question follows: which industry is doing the

identification work? A Bartik point estimate could be driven by one dominant industry that happens to have a large national shock, or it could be a genuine average across many industries with comparable weights. This matters for interpretation and for robustness. Goldsmith-Pinkham et al. (2020) give a clean decomposition.

Proposition 7.3.3 (Rotemberg Decomposition, Proposition 3 in Goldsmith-Pinkham et al., 2020)**.** *The Bartik estimator can be written as*

$$\hat{\beta}_{Bartik} = \sum_k \hat{\alpha}_k \hat{\beta}_k,$$

where

$$\hat{\beta}_k = (Z_k' X^\perp)^{-1} Z_k' Y^\perp$$

is the just-identified IV estimator that uses only the share Z_k of industry k as an instrument, and the weights

$$\hat{\alpha}_k = \frac{g_k Z_k' X^\perp}{\sum_{k'} g_{k'} Z_{k'}' X^\perp}$$

are the Rotemberg weights. *The weights sum to one.*

Each $\hat{\beta}_k$ is what we would get if we used a single share instrument Z_k ($L \times 1$ vector with each row representing the share of industry k in one specific location) to instrument X. The Rotemberg weight $\hat{\alpha}_k$ measures how much industry k's single-instrument IV estimate contributes to the combined Bartik estimate. An industry with a big Rotemberg weight is one whose identification assumption is critical for the overall estimate: if that industry's share is not really exogenous, the Bartik estimate will be badly biased. An industry with a small Rotemberg weight can be misspecified without much damage.

Practical Guidance

Report the industries with the highest Rotemberg weights. Discuss the plausibility of Assumption 7.3.2 for those industries in particular. If the whole identification is really riding on one or two shares, be honest about it and defend those assumptions specifically.

A useful conceptual clarification: the Rotemberg decomposition is different from the GMM equivalence in Proposition 7.3.1. The GMM equivalence says that Bartik in 2SLS with a single combined instrument B equals GMM with K share instruments in a *joint estimation*, using $W = GG'$. The Rotemberg decomposition says that the

Bartik estimator can be rewritten as a weighted average of K *separately estimated single-instrument IV estimators*. Both statements are true and both use the same object, but they describe different aspects of the estimator.

7.3.6 Heterogeneous Treatment Effects: No LATE for Bartik

In the LATE framework of Imbens and Angrist (1994), we saw that a single binary IV with monotonicity identifies the average treatment effect for compliers. Does a Bartik instrument have an analogous LATE interpretation? The answer, developed in Goldsmith-Pinkham et al. (2020), is negative in general. Even in a restricted heterogeneous effect model, the Bartik IV does not have a LATE interpretation. This is not surprising given what we have learned in Chapter 6.

First, for the single share instrument case, in a restricted heterogeneous effect model with a set of monotonicity style assumptions, the Bartik IV can be written as a combination of location level treatment effects with positive weights.

Second, however, for the combined Bartik instrument, monotonicity of each single share instrument is not enough for us to have positive weights. The reason is that different share instruments can push locations in different directions, and combining them can produce non-monotone shifts. Therefore, the weights can be negative. Negative weights make the estimator uninterpretable as any convex average of underlying effects.

You will see the same story unfold in later chapters when we discuss DID with staggered treatment timing and heterogeneous effects. Simple regressions turn out to average local treatment effects in ways that can be pathological when the underlying effects are complicated (heterogeneous, dynamic, and so on), and this is not a coincidence: it is an intrinsic feature of trying to summarize heterogeneous causal effects with a single regression coefficient.

7.3.7 Empirical Suggestions

The GPSS framework leads to several concrete pieces of empirical advice.

Always include location and time fixed effects. As discussed above, initial shares are plausible instruments for outcome *changes*, not levels. In a panel setting, location and time fixed effects deliver the difference specification automatically. In a cross section, use a growth outcome variable rather than a level.

Focus on high Rotemberg weight industries. Report the top industries by weight and defend the exogeneity assumption for each one. If most of the weight is on a single industry, the paper is really identified off that industry, and the discussion should reflect that.

Test 1: Balance on covariates. Suppose there is a location level control d_l that plausibly predicts changes in y through some channel other than x. Test whether the baseline shares are correlated with d_l. Because the shares are supposed to affect y only through x, an economically or statistically significant correlation is a red flag. Goldsmith-Pinkham et al. (2020) suggest also controlling for higher-level aggregate shares (for instance, if industry is defined at the 4-digit level, control for 2-digit shares) to soak up broad category composition.

Test 2: Pre-trends. If the data include a pre-shock period, the specification with baseline shares interacted with pre-period dummies is a DID with growth rate as the shock size. Under the exposure design interpretation, the pre-period coefficient $\tilde{\gamma}_1$ should be zero. Test this for both the overall Bartik IV and for the single-industry instruments with the highest weights.

Test 3: Overidentification. The Bartik-GMM equivalence says that Bartik is an overidentified GMM with shares as instruments. Standard overidentification tests (Hansen's J) can be run at the level of the individual share instruments. As always, a rejection has different potential interpretations: it could be at least one instrument is invalid, or the treatment effect is heterogeneous across instruments, or some other model misspecification. The test cannot distinguish them, so it should be interpreted with care. Goldsmith-Pinkham et al. (2020) do not recommend this test.

7.4 Borusyak, Hull, and Jaravel (2022): Shift as Instrument

The GPSS framework treats shares as the identifying instrument, with the shifts serving as weights. Borusyak et al. (2022) propose the opposite reading: treat the shifts as the identifying instrument, with the shares serving as weights. The identification argument then rests on random assignment of the industry level shocks, not on exogeneity of the location shares.

This is a substantively different framework. It is not that one is right and the other wrong; they capture different empirical situations. When exogeneity plausibly comes from the shares, GPSS is the right framework. When exogeneity plausibly comes from the shifts, BHJ is the right framework.

7.4.1 Setup

BHJ consider a location level regression

$$y_l = \beta x_l + w_l'\gamma + \epsilon_l,$$

where w_l is a vector of controls. The shift-share instrument has the general form

$$z_l = \sum_k s_{lk} g_k, \qquad k = 1, 2, \ldots, K,$$

with s_{lk} the share of category k in location l and g_k the national shift for category k. The moment condition that validates the instrument is

$$E\Big[\sum_l z_l \epsilon_l\Big] = 0.$$

The question is: what does this look like when we plug in the shift-share definition of z_l?

7.4.2 Shock-Level Equivalence

Substitute $z_l = \sum_k s_{lk} g_k$ into the moment condition:

$$E\Big[\sum_l z_l \epsilon_l\Big] = E\Big[\sum_l \sum_k s_{lk} g_k \epsilon_l\Big].$$

Exchange the order of summation and factor out g_k, which does not depend on l:

$$\begin{aligned} E\Big[\sum_l z_l \epsilon_l\Big] &= E\Big[\sum_k \sum_l s_{lk} g_k \epsilon_l\Big] = E\Big[\sum_k g_k \sum_l s_{lk} \epsilon_l\Big] \\ &= E\Big[\sum_k g_k \Big(\frac{\sum_l s_{lk}\epsilon_l}{\sum_l s_{lk}}\Big)\Big(\sum_l s_{lk}\Big)\Big] \\ &= E\Big[\sum_k s_k g_k \bar{\epsilon}_k\Big], \end{aligned}$$

where we have defined

$$s_k = \sum_l s_{lk}, \qquad \bar{\epsilon}_k = \frac{\sum_l s_{lk}\epsilon_l}{\sum_l s_{lk}}.$$

The quantity s_k is the total exposure to category k summed across locations. The quantity $\bar{\epsilon}_k$ is a share-weighted average of the location errors, aggregated up to the shock (category) level; being a ratio, it is invariant to the normalization.

This algebraic manipulation delivers a remarkable observation. The location level moment condition is equivalent to the shock (industry) level moment condition

$$E\Big[\sum_k s_k g_k \bar{\epsilon}_k\Big] = 0.$$

So even though we thought we were running an IV at the location level, the identifying moment is really at the shock level, where the instrument g_k is required to be uncorrelated with a share-weighted aggregate of the location errors. BHJ formalize this as follows.

Proposition 7.4.1 (Shock-Level Equivalence, Proposition 1 in Borusyak et al., 2022)**.** *The shift-share IV estimator $\hat{\beta}$ equals the second stage coefficient of an s_k weighted shock-level IV regression using the shifts g_k as the instrument in estimating*

$$\bar{y}_k = \alpha + \beta \bar{x}_k + \bar{\epsilon}_k,$$

where $\bar{v}_k = \sum_l s_{lk} v_l / \sum_l s_{lk}$ is the exposure-weighted average of a variable v_l across locations.

Remark 7.4.1 (Intercept versus residualized variables). Proposition 7.4.1 is stated with an intercept and with exposure-weighted averages of the raw variables y_l and x_l. Borusyak et al. (2022) state their Proposition 1 with the aggregates of the residu-

alized variables $y_l^{\perp}$ and $x_l^{\perp}$ and no intercept. The two versions coincide under *complete shares*, $\sum_k s_{lk} = 1$ for every location. In that case $\sum_k s_k \bar{v}_k = \sum_l v_l$, so the s_k weighted mean of the aggregates $\bar{v}_k$ is exactly the location mean of v_l, and the intercept in the shock-level regression removes precisely what demeaning at the location level removes. When shares are incomplete, the s_k weighted mean of $\bar{v}_k$ is instead a $\sum_k s_{lk}$ weighted location mean, the two versions differ, and the residualized form (with the sum of shares included as a control) is the one that holds in general. The practice example at the end of this chapter draws shares from a Dirichlet distribution, so shares sum to one within each location and the intercept version applies exactly.

This equivalence has two immediate uses. First, it identifies the true unit of identification. Even though we have L location observations, the identifying variation is really over K industries. If K is small, the effective sample size for the identifying moment is small, regardless of how big L is. Second, it clarifies what exogeneity means. The identifying assumption is that the shocks g_k are as good as randomly assigned, in the sense that they are uncorrelated with the industry level unobservable $\bar{\epsilon}_k$.

7.4.3 Interpreting the Identification Assumption

Let us make the assumption concrete. Suppose we are estimating the effect of a US tariff on employment in China. The tariff shocks g_k vary by industry (steel, textiles, semiconductors, etc.). The location aggregate $\bar{\epsilon}_k$ is a share-weighted average of local unobserved supply shocks, aggregated across locations by how important industry k is in each place.

For the BHJ identification to hold, we need $g_k \perp \bar{\epsilon}_k$: industries that experience a rise in the US tariff should not systematically face different labor supply shocks in the regions where they concentrate. If Chinese steel is intensively produced in Hebei, then a US tariff on steel is a big shock to Hebei. We must assume that Hebei was not simultaneously receiving some other unobserved shock, such as a change in the Gaokao enrollment quota, that would independently affect Hebei's labor market and be correlated with the tariff.

This is a substantive assumption about the sources of variation in g_k. It is often defensible when the shocks are driven by conditions in some other country (US tariffs, from the perspective of Chinese labor markets) or by aggregate technology, but the researcher has to make the case.

7.4.4 Consistency

BHJ establish consistency under two assumptions.

Assumption 7.4.1 (Quasi-random Shock Assignment, Assumption 1 in Borusyak et al., 2022)**.** Conditional on the shares s and the industry level unobservable $\bar{\epsilon}$, the shocks have a common mean:

$$E[g_k \mid \bar{\epsilon}, s] = \mu.$$

Assumption 7.4.2 (Many Uncorrelated Shocks, Assumption 2 in Borusyak et al., 2022)**.** The industry-level Herfindahl index of *normalized* exposure goes to zero,

$$E\Big[\sum_k \tilde{s}_k^2\Big] \to 0, \qquad \tilde{s}_k = \frac{s_k}{\sum_{k'} s_{k'}},$$

and the shocks are pairwise uncorrelated given exposure,

$$\mathrm{Cov}[g_k, g_{k'} \mid \bar{\epsilon}, s] = 0 \text{ for } k \neq k'.$$

The first assumption says the shocks look randomly assigned conditional on the exposures. The second says exposure should not be too concentrated in a few industries, and that shocks are independent across industries. This second assumption is the substantive analogue of $L \to \infty$ in a standard IV setting: we need many effectively independent draws for the LLN to work. In this framework the effective sample size is at the shock (industry) level, not the location level.

Proposition 7.4.2 (Consistency, Proposition 3 in Borusyak et al., 2022)**.** *Under Assumptions 7.4.1 and 7.4.2 plus standard regularity conditions,*

$$\hat{\beta} \xrightarrow{p} \beta.$$

7.4.5 Empirical Suggestions

Three important pieces of practical advice come out of the BHJ framework.

Inference: watch out for standard errors.

Adao et al. (2019) show that traditional (heteroskedasticity-robust or clustered-by-location) standard errors for a shift-share IV are typically wrong. The reason is that

the location level observations are not i.i.d.: the same national shock g_k enters the regression for every location that has any share in industry k, so the errors ϵ_l have a common component and are mechanically correlated across locations with similar industry mixes. Traditional standard errors ignore this correlation and understate uncertainty, sometimes by a lot. BHJ show that the equivalent shock-level regression (Proposition 7.4.1) does not have this problem: at the industry level, the standard errors from the aggregated regression are valid. This gives a clean practical recipe: run the shock level regression and read off the standard error. The Stata package `ssaggregate` automates the aggregation.

Effective sample size and Herfindahl index.

When using the BHJ framework, we must validate Assumption 7.4.2 for consistency. Intuitively, we must have a large number of industries with dispersed exposure, and shocks should not concentrate on a few industries. In practice, we can directly calculate the industry-level Herfindahl index of normalized exposure to make sure it is small enough.

Balance tests.

Location-level balance. Regress a pre-determined location level control r_l (GDP, population, education) on the SSIV z_l. A significant coefficient is a red flag. This can be combined with an Oster-style bound to assess how much unobserved selection would be needed to overturn the estimate. This is similar to the balance test proposed by GPSS.

Location-level balance using an industry-level confounder. Take an industry level covariate r_k, construct the location aggregate $r_l = \sum_k s_{lk} r_k$, and regress that on the SSIV.

Shock-level balance. Aggregate a location level confounder r_l to the industry level via $r_k = \sum_l s_{lk} r_l / \sum_l s_{lk}$, and regress that on the shock g_k. In the BHJ framework this is arguably the most fundamental balance test, because it directly tests the shock-level exogeneity assumption. If shocks g_k are supposed to be quasi-randomly assigned across industries, they should not correlate with any observable industry level characteristic aggregated from location data.

7.5 Comparison of the Two Frameworks

We now have two frameworks that give a shift-share IV a rigorous causal interpretation. They are not competitors so much as different maps of the same territory. Which map to use depends on the terrain of the specific application. Specifically, in one study, there is no need for us to show both shares and shifts are exogenous. As long as the identification assumptions for **either** of these two frameworks are validated, we are safe.

Comparison of GPSS and BHJ

- **GPSS (Goldsmith-Pinkham et al., 2020)**
  - Equivalence: joint GMM using shares as instruments, shifts as weights.
  - Design: exposure DID with shares as differential exposure.
  - Assumption: locations with different shares have parallel trends absent the shock.
- **BHJ (Borusyak et al., 2022)**
  - Equivalence: shock-level regression with shifts as instruments, shares as weights.
  - Design: random assignment of shifts across industries.
  - Assumption: industry level shocks are uncorrelated with industry-aggregated location unobservables.

Under what conditions should we favor one framework over the other?

Prefer GPSS when:

- The exogenous variation is coming from the shares (historical settlement patterns, legacy industrial structure, or some other pre-determined local composition).
- The design is an exposure design in which differential exposure to a common shock is the identifying variation.
- There is a small fixed number of industries, and the sampling is over locations (K fixed, $L \to \infty$).
- The paper focuses on a few specific industries whose shares carry policy meaning.

Prefer BHJ when:

- The exogenous variation is coming from the shifts (industry level shocks that plausibly look random).
- There is a small fixed number of locations, and the sampling is over industries (L fixed, $K \to \infty$).
- The second stage error has a shift-share structure that generates mechanical correlation between the SSIV and the location error, so that inference needs the shock-level correction.

7.6 Application: Autor et al. (2013)

The canonical modern application of Bartik instruments is the China shock paper of Autor et al. (2013). Both Goldsmith-Pinkham et al. (2020) and Borusyak et al. (2022) use it as their running empirical example, precisely because it is representative of how the technique is used in trade and labor economics.

The research question is: what is the effect of the surge in Chinese imports on local labor market outcomes in the United States? The independent variable is a location i level import exposure measure at time t, ΔIPW_{it}, constructed as

$$\Delta\text{IPW}_{it} = \sum_j \frac{L_{ijt}}{L_{jt} \cdot L_{it}} \Delta M_{jt}.$$

where i is commuting zone in the U.S. and j refers to industry. L_{ijt} is the number of workers employed in industry j in commuting zone i. ΔM_{jt} is the change in U.S. imports from China in industry j. The regression outcomes are commuting zone level manufacturing employment, wages, and various welfare and public assistance measures.

The problem for OLS is the reverse causality: a US productivity shock could increase Chinese imports (through demand for cheap inputs) while also affecting local labor markets directly. Autor, Dorn, and Hanson address this with an instrumental variables strategy: instrument ΔIPW_{it} (US) with an analogous exposure variable that uses Chinese imports to eight other high-income countries as the shift. The idea is that Chinese import growth to Germany, Japan, and so on reflects supply-side factors in China (productivity, exchange rates, WTO accession) rather than demand-side factors in the United States. This is a typical Bartik IV.

In the GPSS reading, the identifying instruments are the initial US industry employment shares L_{ijt}/L_{it}, with the national import growth per worker $\Delta M_{jt}/L_{jt}$ as the shift: commuting zones with more manufacturing employment in industries that will later experience large Chinese import growth are more exposed to the shock. The parallel-trend assumption is that, absent the China shock, commuting zones with different industry compositions would have followed parallel trajectories.

In the BHJ reading, the identifying instruments are the industry level Chinese import shocks ΔM_{jt}. The quasi-random assignment assumption is that the industry-by-industry pattern of Chinese import growth (to other high-income countries) is not systematically related to U.S. industry-level unobservables that also affect the aggregated US local labor market outcomes.

Both readings are defensible for the China shock, and both papers show explicitly how their frameworks apply. Students are strongly encouraged to read Autor et al. (2013) together with the discussion sections of Goldsmith-Pinkham et al. (2020) and Borusyak et al. (2022) to see how the same empirical setting is reinterpreted through the two lenses.

7.7 Conclusion

The Bartik or shift-share instrument is a workhorse tool in applied economics, especially in the trade and labor subfields. Its construction is simple: a weighted sum of national industry shocks, with weights given by baseline local industry shares. Its identification, however, requires care.

We have introduced two frameworks that give the shift-share design a rigorous causal interpretation. The first, Goldsmith-Pinkham et al. (2020), treats shares as instruments and shifts as weights. Bartik is equivalent to an over-identified GMM using the K shares. The identifying assumption is exogeneity of the shares, which is plausible for outcome changes but almost never for outcome levels, so specifications must include location and time fixed effects or use a growth outcome. It is essentially one type of exposure design. The Rotemberg decomposition tells us which industries drive the identification and highlights where the exogeneity assumption most matters.

The second framework, Borusyak et al. (2022), treats shifts as instruments and shares as weights. Bartik is equivalent to a shock-level regression using shifts. The identifying

assumption is quasi-random assignment of the industry shocks. Standard errors must account for the shift-share error structure.

The two frameworks are not competitors but complements. A good Bartik paper argues explicitly which framework applies and defends the corresponding assumptions. Applications where exogeneity flows from historical settlement or long-established industrial composition fit the GPSS reading. Applications where the shocks are plausibly random draws from an industry level distribution fit the BHJ reading. The China shock of Autor et al. (2013) is a case in point: both papers use it and both frameworks can be defended.

In the next chapter we turn from IV to difference-in-differences, which shares more with the Bartik exposure design than the surface would suggest. We will see the same themes recur: heterogeneous effects that a simple regression cannot summarize, careful design-based reasoning about where identifying variation actually comes from, and modern econometric tools that clean up the messiness.

Homework Problems

1. **Verifying the Bartik-GMM Equivalence.**

   Consider the special case of $K = 2$ industries and one time period with no controls other than a constant, so $X^\perp$ and $Y^\perp$ are the demeaned X and Y, with entries $x_l - \bar{x}$ and $y_l - \bar{y}$. Suppose there are L locations with observed data $(y_l, x_l, z_{l1}, z_{l2})$ and national shifts (g_1, g_2). Let $B_l = z_{l1}g_1 + z_{l2}g_2$.

   (a) Write out the Bartik estimator $\hat{\beta}_{\text{Bartik}} = B'Y^\perp / B'X^\perp$ in summation notation using the demeaned variables.

   (b) Write out the GMM estimator with weight matrix $W = GG'$ where $G = (g_1, g_2)'$, using $X^\perp$ and $Y^\perp$. Simplify.

   (c) Show algebraically that the two coincide.

   (d) Now substitute the constraint $z_{l2} = 1 - z_{l1}$. Show that the Bartik estimator reduces to a just-identified IV estimator using z_{l1} alone.

2. **Rotemberg Weights and Interpretation.**

   Consider a Bartik instrument with $K = 3$ industries. Suppose the single-industry

IV estimates and unnormalized weights are:

Industry k	$\hat{\beta}_k$	Unnormalized weight $g_k Z_k' X^{\perp}$
1	2.0	0.6
2	1.0	0.3
3	-3.0	0.1

(a) Compute the normalized Rotemberg weights $\hat{\alpha}_k$ and verify they sum to one.

(b) Compute the combined Bartik estimator $\hat{\beta}_{\text{Bartik}} = \sum_k \hat{\alpha}_k \hat{\beta}_k$.

(c) Suppose the exogeneity assumption is violated for industry 1 but holds for industries 2 and 3. Which industry's misspecification would do the most damage to the Bartik estimate, and why? Which industries could be misspecified with relatively little damage?

(d) Now suppose we relabel the industries so that industry 3 is called industry 1 (swap the labels). Do the Rotemberg weights and $\hat{\beta}_{\text{Bartik}}$ change?

3. **Deriving the Shock-Level Equivalence.**

Consider the BHJ setup with one period, no controls, and a shift-share instrument $z_l = \sum_k s_{lk} g_k$. Assume $\sum_l s_{lk} = s_k$ for each k.

(a) Show that the location-level moment condition $E[\sum_l z_l \epsilon_l] = 0$ can be rewritten as the shock-level moment condition $E[\sum_k s_k g_k \bar{\epsilon}_k] = 0$, where $\bar{\epsilon}_k = \sum_l s_{lk} \epsilon_l / \sum_l s_{lk}$.

(b) Explain in words the difference between the location level exogeneity assumption $E[z_l \epsilon_l] = 0$ and the shock level exogeneity assumption $E[g_k \bar{\epsilon}_k] = 0$. Which of the two is easier to defend in the context of the China shock?

(c) Suppose one industry accounts for 60% of total exposure, $\tilde{s}_1 = 0.6$, while the remaining $K - 1$ industries share the remaining 40% equally. Discuss what this concentration does to Assumption 7.4.2 and to the effective sample size for identification.

Practice Example: Bartik IV with Simulated Shift-Share Data

In this practice example we simulate a shift-share environment, verify that OLS is biased, verify that the Bartik instrument recovers the true structural coefficient, and compute the Rotemberg decomposition. The setup follows the GPSS framework, but the equivalent shock-level regression from BHJ is computed at the end as a cross check.

Data generating process. Consider $L = 300$ locations and $K = 6$ industries. Baseline shares z_{lk0} are drawn from a Dirichlet distribution so they sum to one for each location. National industry shifts g_k are measured in percentage points and fixed at

$$(g_1, g_2, g_3, g_4, g_5, g_6) = (15, -3, 2, 8, -5, 3),$$

so industry 1 has a large positive shock, industry 4 a moderate positive shock, and the rest are small or negative. Measuring the shifts (and hence x and y) in percentage points is not cosmetic: because B_l is a convex combination of the g_k, writing the same values as decimal fractions would put the variance of the instrument more than three orders of magnitude below the noise variance in x, the first stage F statistic would fall below one, and no IV method could work. A local unobserved shock $\epsilon_l \sim \mathcal{N}(0, 1)$ is drawn independently of everything else. The endogenous variable is

$$x_l = \sum_k z_{lk0} g_k + \lambda \epsilon_l + \eta_l,$$

where $\lambda = 0.8$ generates the endogeneity and $\eta_l \sim \mathcal{N}(0, 0.5)$ is a pure first-stage shock. The outcome is

$$y_l = \alpha + \beta_0 x_l + \epsilon_l,$$

with $\beta_0 = 1.0$ the true parameter.

Predicted results. OLS regresses y_l on x_l and picks up $\beta_0 + \lambda \operatorname{Var}(\epsilon_l)/\operatorname{Var}(x_l)$, which for these parameters is roughly $1 + 0.8/4.6 \approx 1.17$. Bartik IV using $B_l = \sum_k z_{lk0} g_k$ should recover $\beta_0 \approx 1$. Rotemberg weights should be dominated by industry 1 (largest positive shift), with moderate weights on industries 4 and 5 (industry 5 pairs a negative shift with a negative share covariance, which produces a positive weight) and small weights, possibly negative, on the rest. The BHJ shock-level regression, with six

observations, should give an identical point estimate to the location-level Bartik IV.

Analysis steps. The do-file (`src/chapter 7/ch7_bartik.do`) performs:

1. Simulate the shift-share DGP.
2. Descriptive statistics: verify shares sum to one, tabulate the fixed shifts, summarize x and y.
3. OLS of y on x (biased).
4. Bartik IV: construct B_l, run 2SLS.
5. Single-industry IV: for each k, run IV using z_{lk0} alone; collect $\hat{\beta}_k$.
6. Rotemberg weights: compute $\hat{\alpha}_k$ and verify $\sum_k \hat{\alpha}_k \hat{\beta}_k = \hat{\beta}_{\text{Bartik}}$.
7. BHJ shock-level equivalence: aggregate y and x to the industry level via share-weighted averages, run the shock-level IV using g_k as the instrument, and verify the coefficient equals the Bartik estimate.
8. Key lessons summary.

Key lessons.

- OLS is biased when the endogenous variable is correlated with the outcome error. The direction and magnitude of the bias depend on the sign and size of λ.
- The Bartik instrument B_l is exogenous by construction, because both the baseline shares and the national shifts are drawn independently of ϵ_l. Bartik IV recovers β_0.
- The Rotemberg decomposition reveals which industries drive the estimate. Industries with large shifts and shares that are systematically correlated with x get large weights; industries with small shifts or shares uncorrelated with x get small weights.
- The BHJ shock-level equivalence gives the same point estimate as the location-level Bartik IV, confirming Proposition 7.4.1. In real applications this equivalence delivers valid standard errors via `ssaggregate`, which the location level regression cannot.

The do-file is located at `src/chapter 7/ch7_bartik.do` and the simulated dataset at `data/chapter 7/ch7_bartik.dta`.

Part III

Panel Data and Discontinuities

Chapter 8

Causal Inference with Panel Data I

8.1 Introduction

In the previous chapters we mostly worked with cross-sectional data (except for Chapter 7): one observation per unit, all measured at a single point in time. This chapter adds a new dimension. When we observe the same units repeatedly over time, we have *panel data*, and a whole new source of identifying variation opens up: variation across time *within* the same individual, firm, city, or country. Panel data let us compare a unit to itself, which eliminates every confounder that stays constant over time.

This chapter develops the panel data toolkit for causal inference in four steps. First, we introduce the fixed effect model and its three equivalent or nearly equivalent estimators: the fixed effect estimator, the dummy variable estimator, and the first difference estimator. Second, we combine unit and time fixed effects into the two way fixed effect (TWFE) model and study its most famous special case, the difference-in-differences (DID) design, using the minimum wage study of Card and Krueger (1994) as the running example. The key identifying assumption is parallel trends, and we discuss the traditional ways of assessing it: plotting raw trends and running event study regressions. Along the way we emphasize a question that every applied researcher must be able to answer: *which variation in the data identifies the parameter?*

Third, we present three modern extensions. The recent literature has taught us to be much more careful about pre-trend testing: (1) Roth (2022) shows that conventional event study pre-tests often have low power, and that pre-testing itself distorts inference.

(2) When only a few units are treated with no clean control group, the synthetic control method from Abadie et al. (2010) constructs an artificial control unit as a weighted combination of untreated units. (3) When a second within-state comparison group is available, triple differences (DDD) can difference out non-parallel trends.

Fourth, a preview: the next chapter continues with panel data and treats staggered adoption designs, where treatment turns on at different times for different units and the TWFE regression can behave badly. The tools of this chapter are the foundation for that discussion.

8.2 Fixed Effects

8.2.1 Setup

Consider a classic question: what is the impact of military service on wages? Following the fixed effects framework in Chapter 5 of Angrist and Pischke (2009), let i index individuals and t index time. We observe wage Y_{it}, military service status D_{it}, and time varying covariates X_{it}. The worry is an unobserved individual attribute, ability A_i, that affects both the decision to serve and later wages. Assume a constant treatment effect and a linear CEF:

$$\begin{aligned} Y_{it} &= \alpha + \rho D_{it} + A_i'\gamma + X_{it}'\beta + \epsilon_{it}, \\ E[Y_{it} \mid A_i, X_{it}, D_{it}] &= \alpha + \rho D_{it} + A_i'\gamma + X_{it}'\beta. \end{aligned} \tag{8.1}$$

The unobserved confounder A_i has two crucial properties: it is *unobserved*, so we cannot control for it directly, and it is *time invariant*, which is what panel data will let us exploit. The identifying assumption is that selection into treatment is fully captured by A_i and X_{it}:

$$\epsilon_{it} \perp\!\!\!\perp D_{it} \mid A_i, X_{it}.$$

If we could observe A_i, we would simply control for it. We cannot, but with panel data we do not need to. There are three simple ways to estimate ρ.

8.2.2 Three Estimators

Method 1: The fixed effect (within) estimator. The FE estimator is a deviation from means estimator. Take individual level time averages of both sides of (8.1):

$$\bar{Y}_i = \alpha + \rho \bar{D}_i + A_i'\gamma + \bar{X}_i'\beta + \bar{\epsilon}_i,$$

where $\bar{Y}_i = \frac{1}{T}\sum_t Y_{it}$ and so on. Now subtract this from the original equation:

$$\begin{aligned} Y_{it} - \bar{Y}_i &= \rho\,(D_{it} - \bar{D}_i) + (A_i'\gamma - A_i'\gamma) + (X_{it} - \bar{X}_i)'\beta + (\epsilon_{it} - \bar{\epsilon}_i) \\ &= \rho\,(D_{it} - \bar{D}_i) + (X_{it} - \bar{X}_i)'\beta + (\epsilon_{it} - \bar{\epsilon}_i). \end{aligned} \tag{8.2}$$

The unobserved time invariant term $A_i'\gamma$ is identical in every period for individual i, so it equals its own time average and cancels. Running OLS on the demeaned equation (8.2) delivers ρ without ever observing A_i.

Method 2: The dummy variable estimator. Alternatively, saturate the individual dimension with a full set of individual dummies. Collect the constant and the individual specific term into one intercept per person:

$$Y_{it} = \alpha_i + \rho D_{it} + X_{it}'\beta + \epsilon_{it}, \qquad \alpha_i \equiv \alpha + A_i'\gamma. \tag{8.3}$$

The unobserved A_i is absorbed into the estimated dummy coefficient α_i. Running OLS on (8.3) with N individual dummies again delivers ρ. Because OLS on group indicators just fits the group mean, then By the Frisch–Waugh–Lovell logic, regressing on a full set of individual dummies is numerically identical to demeaning within individuals. Thus, the dummy estimator and the FE estimator produce exactly the same point estimates, standard errors, and other main statistics. For this reason the two are used interchangeably and jointly called the "FE model."

Method 3: The first difference (FD) estimator. A third route removes A_i by differencing across adjacent periods. Let $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$. Subtracting the period $t-1$ equation from the period t equation:

$$\Delta Y_{it} = \rho\,\Delta D_{it} + \Delta X_{it}'\beta + \Delta\epsilon_{it}. \tag{8.4}$$

Again the time invariant A_i drops out, this time through differencing rather than

demeaning.

8.2.3 Comparing FE, Dummy, and FD

All three estimators exploit the same source of variation: changes over time within the same person. But they are not always numerically identical, and the differences matter in practice.

FE, Dummy, and FD: How They Relate

- **FE and Dummy are always identical.** Point estimates, standard errors, and the main regression statistics coincide. We refer to both as the FE model.
- **FE and FD are identical when $T = 2$.** With two periods, the within deviation and the first difference are the same transformation up to scale.
- **FE and FD differ when $T > 2$,** and the choice between them is an inference question about the error process:
  - If ϵ_{it} are serially *uncorrelated* shocks, FE is more efficient. First differencing manufactures serial correlation in $\Delta\epsilon_{it}$ (adjacent differences share a common ϵ term), which FD then has to live with. Serial correlation invalidates the usual statistical inference based on the i.i.d. assumption and damages efficiency.
  - If ϵ_{it} follows a *random walk*, that is, $\epsilon_{it} = \epsilon_{i,t-1} + u_{it}$ with u_{it} i.i.d., then FD is better, because the original errors are serially correlated but $\Delta\epsilon_{it} = u_{it}$ is not.

8.3 Difference-in-Differences and TWFE

8.3.1 The Two Way Fixed Effect Model

With panel data we can usually control for both individual and time fixed effects:

$$Y_{it} = \rho D_{it} + X_{it}'\beta + \lambda_t + \alpha_i + \epsilon_{it}. \tag{8.5}$$

This is the **two way fixed effect (TWFE) model**. The individual effects α_i absorb everything constant within a unit, and the time effects λ_t absorb everything common to all units in a period (macro shocks, seasonality, national policy).

The **difference-in-differences (DID)** design is a special case of the TWFE model. In a typical DID application, some policy is implemented at a higher level of aggregation than the individual, such as a province or a city, and D_{it} is a binary indicator for whether individual i at time t is covered by the policy. We control for unit (or province, or city) fixed effects and time fixed effects, and read the policy effect off ρ.

8.3.2 The Card and Krueger Setting

The canonical DID example is Card and Krueger (1994) on the minimum wage. On April 1, 1992, New Jersey raised its state minimum wage. In neighboring Pennsylvania, nothing happened. Card and Krueger collected employment data from fast food restaurants in New Jersey and Pennsylvania in February 1992 (before the change) and November 1992 (after the change). The question is whether the minimum wage increase reduced employment, as a simple competitive model of the labor market predicts.

For restaurant i in state s at time t, denote employment by Y_{ist} and the policy dummy by D_{st}. In this two state, two period setting, $D_{st} = NJ_s \cdot d_t$, where NJ_s indicates New Jersey and d_t indicates the post policy period. Our target is the average treatment effect on the treated,

$$E[Y_{1ist} - Y_{0ist} \mid D_{st} = 1].$$

The problem is the usual missing counterfactual: after the policy, we observe only the treated outcome Y_{1ist} for New Jersey restaurants. How would employment in New Jersey have evolved *without* the policy?

8.3.3 The Parallel Trend Assumption

The DID answer: use the untreated state as the control group, and assume the treated state would have followed the same time path as the control state in the absence of the policy.

Assumption 8.3.1 (Parallel Trends)**.** There are no differential trends across treated and untreated states absent the policy change:

$$E[Y_{0ist} \mid s, t] = \gamma_s + \lambda_t. \tag{8.6}$$

The untreated potential outcome decomposes into a state component and a time component, with no interaction: no term like η_{st} appears in $E[Y_{0ist} \mid s, t]$.

Assumption 8.3.1 is additive separability of the untreated outcome in state and time. States can differ in levels (γ_s), and time can shift all states together (λ_t), but there is no state specific trend. This is what makes Pennsylvania an acceptable stand-in for the New Jersey counterfactual.

8.3.4 Identification

Under parallel trends, the policy effect δ is identified by the regression

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \epsilon_{ist}. \tag{8.7}$$

To see why, take two differences. The *first difference* compares the same state across time:

$$E[Y_{ist} \mid s = PA, t = Nov] - E[Y_{ist} \mid s = PA, t = Feb] = \lambda_{Nov} - \lambda_{Feb}, \tag{8.8}$$

$$E[Y_{ist} \mid s = NJ, t = Nov] - E[Y_{ist} \mid s = NJ, t = Feb] = \lambda_{Nov} - \lambda_{Feb} + \delta. \tag{8.9}$$

Each within state change nets out the state level γ_s but still contains the common time shift. The *second difference* compares the two state level changes:

$$(8.9) - (8.8) = \delta. \tag{8.10}$$

The common time trend cancels, and what remains is the policy effect. We difference once over time and once over states, hence the name difference-in-differences. The logic is drawn in Figure 8.1: the control group's observed trend is used to construct the treated group's counterfactual, and the treatment effect is the gap between the observed and counterfactual treated outcomes after the policy.

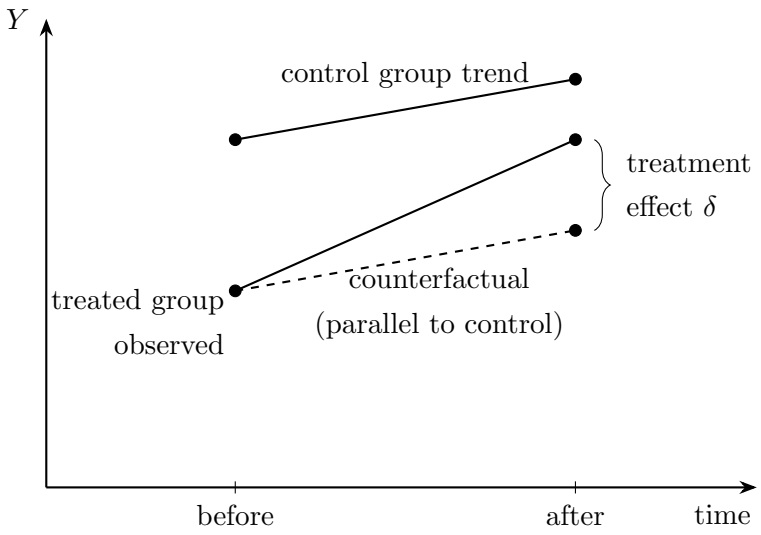


Figure 8.1: The DID identification logic. The treated group's counterfactual path (dashed) is constructed by shifting the control group's trend down to the treated group's pre-period level. The treatment effect is the post-period gap between the observed and counterfactual treated outcomes.

8.3.5 Testing the Parallel Trend Assumption

After the policy is implemented at time t_0, we can no longer observe Y_0 for the treated group, so the parallel trend assumption is intrinsically untestable in the post period. What we *can* examine is the pre period: did the two groups trend in parallel before t_0? This is the **pre-trend test**, and there are two simple ways to do it.

Method 1: Draw the picture. Plot the average outcome for treated and control groups over time and look at the pre-policy period. In the Card and Krueger follow-up data (Card and Krueger, 2000), employment in New Jersey and Pennsylvania fast food restaurants tracks each other reasonably well before the policy dates, which supports the design. In other applications the picture can be damning: when the treated group is already on a visibly different trajectory before the policy (rising while the control is flat, say), the parallel trend assumption loses credibility no matter what the post period shows. Always draw this picture first. It is the cheapest and most honest diagnostic available.

Method 2: The event study regression. Suppose we have data from $-T$ to T' in event time, with the policy implemented at $t = 0$ for the treated group. Let D_s be the

treated group dummy. Run

$$Y_{ist} = \gamma_s + \lambda_t + \sum_{\tau \neq -1} \mathbf{1}(t = \tau)\, \delta_\tau D_s + \epsilon_{ist}. \tag{8.11}$$

Each event time τ gets its own coefficient δ_τ, which measures the difference in trends between treated and untreated groups at τ, relative to the omitted baseline. The period just before the policy, $\tau = -1$, is conventionally omitted as the baseline. Note the contrast with the static DID: there we split time into just two bins, before and after, with $D_{st} = D_s d_t$ switching on only for the treated group after implementation; in the event study we give every time point its own parameter.

The estimated δ_τ are then plotted against event time with confidence intervals, as in Figure 8.2. The reading is:

- Points *before* $t = 0$ insignificant and close to zero $\Rightarrow$ the pre-trend is parallel.
- Points *after* $t = 0$ show the dynamic policy effect, period by period.

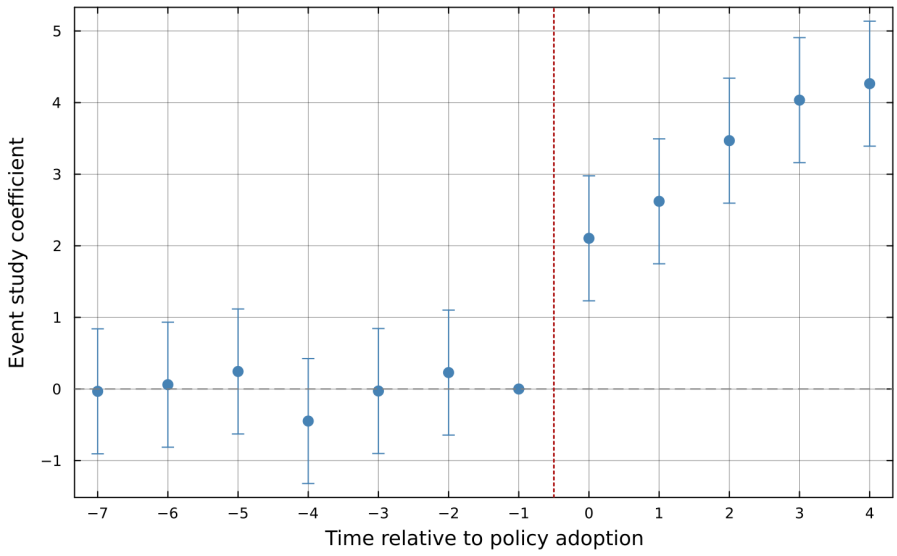


Figure 8.2: An event study coefficient plot from a simulated DID panel. Each point is $\hat{\delta}_\tau$ with a 95% confidence interval; the period $\tau = -1$ is the omitted baseline (normalized to zero, shown without an interval). Pre-period coefficients are small and insignificant, consistent with parallel trends; post-period coefficients trace out a dynamic treatment effect.

A Traditional DID Procedure

1. Draw the changes of Y over time for treated and control groups as descriptive evidence.
2. Run the main DID regression.
3. Run the event study regression to check the pre-trend and to trace the dynamic effect.
4. Remember to cluster your standard errors. (More details on clustering in Chapter 11.)

8.4 Identification Source: Which Variation Are You Using?

When you are doing causal research, a central question is: *what kind of variation is used to identify the causal effect?* This question is very, very important. It determines how you can interpret your results, it determines which assumption you are actually invoking, and it therefore determines how you should defend your research design.

With cross-sectional data the answer is usually easy. With panel data it becomes complicated, because there are more dimensions along which variation lives. Applied papers control for many fixed effects at different levels, sometimes combined with IV, RD, or other designs. Whatever the combination, you should always be able to state precisely where the identifying variation comes from.

Consider a simple example: the effect of working experience on wages, for individual i from family j at time t:

$$wage_{ijt} = \beta_0 + \beta_1 exp_{ijt} + \epsilon_{ijt}. \tag{8.12}$$

The fixed effects you include determine the comparison the regression makes.

Fixed Effects and the Variation They Leave

For regression (8.12):

- **Time FE only:** you are using variation across individuals and families (i, j level) within the same year.

- **Individual FE only:** you are using variation across time (t level) for the same person.
- **Family FE and time FE:** you are using variation across individuals within the same family in the same year ($i \mid j$ level).
- **Individual FE and time FE:** you are using variation in time trends across different people ($i \times t$ level).

Each row of the box corresponds to a different comparison, a different exogeneity requirement, and a different economic interpretation. Two researchers running "the same" regression with different fixed effects are, in identification terms, running different research designs.

8.5 Pre-Trend Testing Reconsidered

We now turn to the first of three modern extensions of DID: a more careful look at pre-trend testing. Is the event study a perfect tool for testing parallel trends? It is good, but far from perfect. The key reference is Roth (2022), which examines what event study pre-tests actually deliver in practice.

8.5.1 Problem 1: Low Statistical Power

The first problem is that pre-trend tests frequently have low power: they are likely to commit type II errors, failing to detect violations that are large enough to matter. Pre-existing trends that produce meaningful bias in the DID estimate may simply not be detected by the test.

To make this concrete, suppose parallel trends is violated linearly: the differential trend between treated and control groups is $\delta_{1t} - \delta_{0t} = \gamma t$ for some slope γ. Roth (2022) runs Monte Carlo simulations calibrated to data from a large set of published papers and asks: how big must the violation be before the standard pre-test detects it reliably? The answer is sobering. For the test to detect the violation 80% of the time, the implied bias in the treatment effect estimate has to be roughly as large as the estimated treatment effect itself. In other words, a pre-existing trend big enough to explain the entire headline result can still slip past the test most of the time. The bias has to be very large for you to detect it.

Why does this happen? It goes back to the nature of statistical testing. Recall the two error types:

- **Type I error:** H_0 is true but we reject it. Its probability is the significance level α.
- **Type II error:** H_0 is false but we fail to reject it. Its probability is β.
- **Power:** the probability of rejecting H_0 when it is false, $1 - \beta$.

The four possible outcomes of a hypothesis test are summarized in Figure 8.3.

Null hypothesis is ...	**True**	**False**
Rejected	**Type I error** False positive Probability $= \alpha$	**Correct decision** True positive Probability $= 1 - \beta$
Not rejected	**Correct decision** True negative Probability $= 1 - \alpha$	**Type II error** False negative Probability $= \beta$

Figure 8.3: Type I and type II errors: the four possible outcomes of a hypothesis test, with the probability of each outcome.

There is an unavoidable trade-off. When you choose a critical value, moving it in one direction raises α, and moving it in the other raises β. You can decrease the type I error rate only by increasing the type II error rate, and vice versa. If you want H_0 to be rejected less easily, you must tolerate a larger probability of a false negative.

Figure 8.4 visualizes the trade-off. The critical value splits the horizontal axis between the two sampling distributions: the right tail of the null distribution beyond the critical value is the type I error rate α, and the left tail of the alternative distribution below it is the type II error rate β. Pushing the critical value to the right means you are setting a higher bar to reject the null hypothesis. It shrinks α but inflates β. Pulling it to the left does the reverse. There is no choice that shrinks both.

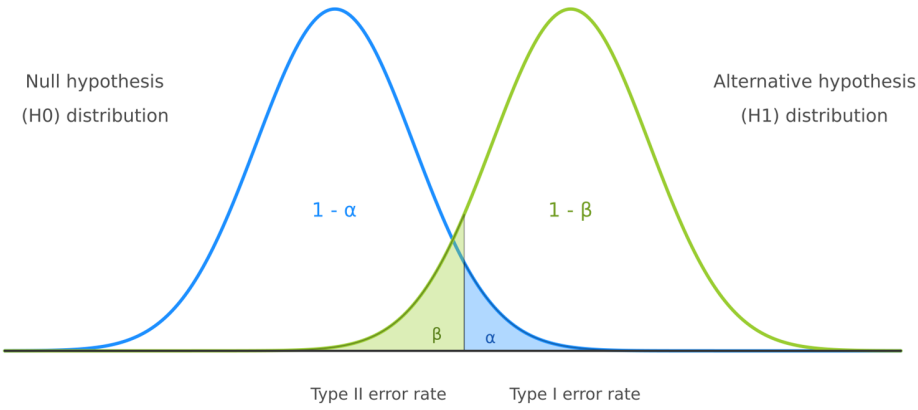


Figure 8.4: The probability of making type I and type II errors. The null sampling distribution (blue) and the alternative (green) overlap; at the critical value, the shaded right tail of the null is α and the shaded left tail of the alternative is β. Moving the critical value trades one error rate against the other.

Here is the catch. Traditional hypothesis testing is deliberately conservative about *rejecting* H_0: we cap α at 10%, 5%, or 1%, and accept whatever β results. That convention is designed for situations where a false rejection is the costly mistake. But in pre-trend testing the roles are reversed. The null hypothesis is "trends are parallel," and the researcher is hoping *not* to reject it. What we should be conservative about is *failing to reject* a false null, that is, we care about power. A testing convention that makes rejection very hard mechanically makes it very likely that your design passes the pre-test, whether or not parallel trends actually holds.

8.5.2 Problem 2: Pre-Testing Distorts Inference

The second problem is subtler. In practice, researchers run the pre-test and proceed only if the design passes. This turns the pre-test into a sample selection device operating across research designs, and conditioning on passing has consequences:

2a. **Variance understatement.** Conditional on having passed a pre-test, the usual standard errors understate the true sampling variability of the estimator.

2b. **Bias exacerbation.** If parallel trends is in fact violated, conditioning on passing the event study test can make the bias of the point estimate *worse*, not

better. The draws in which the violation happens to look small in the pre-period, which are exactly the draws that pass, are systematically unrepresentative. Roth (2022) shows the bias is certainly exacerbated in common configurations, such as monotone trend violations with homoskedastic errors.

The overall effect of pre-trend testing is therefore ambiguous: it screens out some non-parallel designs, which is good, but among designs with a real violation that survive the screen, it makes the surviving estimates more biased, which is bad.

8.5.3 Practical Suggestions

Practical Suggestions Following Roth (2022)

- **Most important: always use your economic knowledge to assess the parallel trend assumption.** The pre-test is a diagnostic, not a substitute for thinking about why the treated and control groups should trend together.
- Do not conclude you are safe just because the event study does not detect any non-parallel pre-trend.
- Do power calculations against economically relevant violations of parallel trends.
- If you know the functional form of the differential trend, control for it directly.
- Conduct sensitivity analysis using Rambachan and Roth (2023): compute bounds on your estimates under a specified amount of parallel trend violation, and report how large a violation would be needed to overturn the conclusion.

8.6 Synthetic Control

8.6.1 Main Idea

In the previous analysis we considered micro-level units like individuals and restaurants. In a large micro-level dataset, it is not hard to find comparable treated and control groups. However, what if the treated unit is only one country or one province? It may be hard to find another proper untreated country or province that is similar enough. What should we do? Just create one synthetic control unit. This is the synthetic control method, developed in Abadie et al. (2010, 2015) and reviewed in Abadie (2021). At heart it is a matching method applied to aggregate units.

When the units of observation are a small number of aggregate entities, such as states or countries, a combination of unaffected units often provides a more appropriate comparison than any single unaffected unit alone. Rather than picking one control state, we construct a weighted average of many control states that mimics the treated state as closely as possible before the treatment.

The canonical application is Abadie et al. (2010) on California's Proposition 99, a large scale tobacco control program implemented in 1988 that combined a 25 cent per pack excise tax with restrictions such as a ban on cigarette vending machines in public areas accessible to juveniles. The empirical problem is that California's cigarette sales were not trending in parallel with other states before 1988, and even the average of all other states tracks California poorly. No single state, and not the national average, is a credible control. The solution: combine the other states into a man-made "synthetic California" whose pre-1988 sales path matches the real California almost exactly. After 1988, the gap that opens between real and synthetic California estimates the effect of Proposition 99.

8.6.2 Formal Setting

Suppose we have units $j = 1, 2, \ldots, J+1$ (provinces, states, countries) observed over T periods. Periods $1, \ldots, T_0$ are pre-treatment and treatment starts in period $T_0 + 1$; $j = 1$ is the treated unit. The remaining units $j = 2, \ldots, J+1$ are called the **donor pool**: the synthetic control will be built from them. For each unit we observe a vector of characteristics X_{kj}, which may include pre-treatment values of the outcome itself.

The Predictors Must Be Unaffected by Treatment

The characteristics X_{kj} used to construct the synthetic control must be unaffected by the treatment. Pre-treatment outcomes can be included; post-treatment outcomes must never be.

Define potential outcomes Y_{jt}^I (with intervention) and Y_{jt}^N (without intervention). The treatment effect of interest for the treated unit is

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N \quad \text{for } t > T_0,$$

which can vary across time.

Definition 8.6.1 (Synthetic Control Estimator)**.** A synthetic control is a weighted average of the units in the donor pool. Given weights $w_j \geq 0$ for $j = 2, \ldots, J+1$ with $\sum_{j=2}^{J+1} w_j = 1$, the synthetic control estimate of the treated unit's untreated outcome is

$$\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt},$$

and the estimated treatment effect is

$$\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^N.$$

8.6.3 Choosing the Weights

How are the weights determined? Stack the characteristics of the treated unit into X_1 and those of the donors into X_0, with the predictor variables indexed by $h = 1, \ldots, k$. The unit weights $W = (w_2, \ldots, w_{J+1})'$ are chosen to minimize the weighted Euclidean distance

$$\|X_1 - X_0 W\| = \Big(\sum_{h=1}^{k} v_h \left(X_{h1} - w_2 X_{h2} - \cdots - w_{J+1} X_{h,J+1} \right)^2 \Big)^{1/2}. \tag{8.13}$$

We are searching for the combination of donors that mimics our treated unit the best, predictor by predictor. Watch out for the two different sets of weights in (8.13):

- w_j is the weight assigned to each *unit* (state) in forming the synthetic control.
- v_h is the weight assigned to each *characteristic* (predictor) when we compute the distance that determines w. It encodes how important each predictor is for the match.

A concrete illustration comes from the German reunification application of Abadie et al. (2015): how would West German GDP have evolved without the 1990 reunification? The treated unit is West Germany; the donor pool is a set of other OECD countries. The predictor weights v are attached to economic growth predictors such as GDP per capita, trade openness, inflation, industry share, population education level, and the investment rate. The resulting synthetic West Germany matches the real West Germany's pre-1990 predictor means far better than the OECD average or any single country. The country weights w turn out to concentrate on a handful of donors (in

the published application: Austria, the United States, Japan, Switzerland, and the Netherlands, with all other weights at zero), and the synthetic Germany tracks real West German GDP per capita almost perfectly for three decades before reunification.

8.6.4 Choosing v by Cross-Validation

The predictor weights v are a tuning parameter, and they are chosen by a training and validation split of the pre-treatment data:

1. **Divide** the pre-treatment sample into two parts: a training period $t = 1, \dots, t_0$ and a validation period $t = t_0 + 1, \dots, T_0$.
2. **For every candidate** V, compute the unit weights $\tilde{w}(V)$ by minimizing (8.13) on the training data.
3. **Select** V^* as the candidate that minimizes the mean squared prediction error on the validation data:

$$\frac{1}{T_0 - t_0} \sum_{t=t_0+1}^{T_0} \Big(Y_{1t} - \tilde{w}_2(V) Y_{2t} - \cdots - \tilde{w}_{J+1}(V) Y_{J+1,t} \Big)^2 .$$

4. **Recompute** the final unit weights $W^* = \tilde{w}(V^*)$ on the validation period data.

This is exactly the logic of model selection by cross-validation from Chapter 3, transplanted into the comparative case study world: fit on one part of the data, evaluate predictive performance on another, and pick the tuning parameter that predicts best out of sample.

Figure 8.5 illustrates the method on a simulated example: the synthetic control tracks the treated unit closely through the pre-period, then a gap opens after the treatment date, and that gap is the estimated effect.

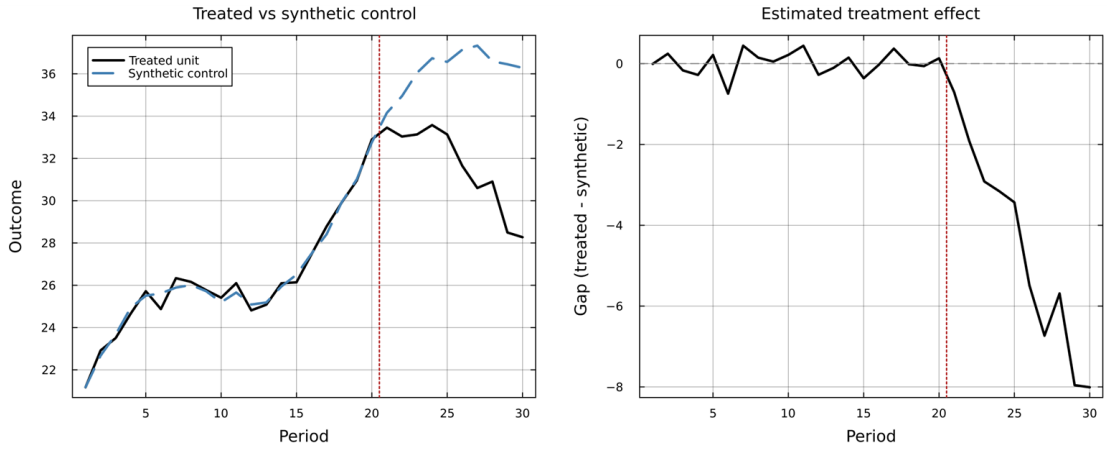


Figure 8.5: Synthetic control in a simulated example. Left: outcome paths of the treated unit and its synthetic control, with the treatment date marked by the dotted vertical line. The synthetic control matches the treated unit closely before treatment. Right: the gap between treated and synthetic outcomes, which estimates the (here negative and widening) treatment effect.

Things to Remember for Synthetic Control

- Post-treatment outcomes cannot be used in X. Pre-treatment outcomes can.
- There can be bias if there are unobserved endogenous characteristics that cannot be matched.
- When the distance $\|X_1 - X_0 W^*\|$ is large, the pre-treatment fit is bad: do not use synthetic control in that case.
- If the donor pool is too large, there can be over-fitting. Choose the donor pool judiciously.
- All unit weights must be nonnegative and sum to one, so each w_j lies within $[0, 1]$.

8.7 Triple Differences (DDD)

8.7.1 Main Idea

Another method to extend DID is triple differences (DDD). Return to Card and Krueger (1994) and suppose we worry that Pennsylvania and New Jersey have genuinely different time trends in average wages, driven by some common shock that hits the two states differently. Then the simple two state DID is contaminated.

But notice what the minimum wage does and does not touch. A minimum wage increase affects the cashier at McDonald's, whose wage sits at the minimum, but it does not affect teachers, whose wages are far above it. If whatever differential state level shock we worry about moves cashiers and teachers alike, then the *gap* between cashier and teacher wages has the same trend in Pennsylvania and New Jersey absent the policy. The only genuinely treated group is cashiers in New Jersey after the policy.

8.7.2 The Three Way Fixed Effect Model

Denote i as the individual, g as the occupation group ($g = 1$ for cashiers), s as the state ($s = 1$ for New Jersey), and t as time ($t = 1$ for November, the post period). Extend the TWFE model to

$$Y_{it} = \rho D_{it} + \lambda_{g,t} + \delta_{s,t} + \alpha_i + \epsilon_{it}. \tag{8.14}$$

This is a **three way fixed effect (3WFE) model**. We control for individual level fixed effects α_i (which absorb group and state levels, since occupation and state are fixed within individual), occupation by time effects $\lambda_{g,t}$, and state by time effects $\delta_{s,t}$. The treatment dummy switches on only for the triple intersection: $D_{it} = 1$ only if $t = 1$, $g = 1$, and $s = 1$.

The state by time effects $\delta_{s,t}$ are exactly what the two state DID could not control: they allow New Jersey and Pennsylvania to follow different aggregate trends. The occupation by time effects $\lambda_{g,t}$ allow cashiers and teachers nationwide to trend differently. What must be excluded is a shock specific to the triple cell (cashiers in New Jersey after the policy) other than the policy itself.

8.7.3 Identification

The extended parallel trend assumption is that the error is orthogonal to the treatment given the fixed effects, $E[\epsilon_{it} \mid \lambda_{g,t}, \delta_{s,t}, \alpha_i, D_{it}] = 0$. Under this assumption, the policy effect ρ is recovered by three nested differences. Write $E[Y_{ist} \mid s, g, t]$ for the cell means implied by (8.14).

First difference (same state, same occupation, across time; the individual effect α_i

cancels):

$$E[Y \mid s{=}0, g{=}0, t{=}1] - E[Y \mid s{=}0, g{=}0, t{=}0] = \lambda_{0,1} - \lambda_{0,0} + \delta_{0,1} - \delta_{0,0}, \tag{8.15}$$
$$E[Y \mid s{=}0, g{=}1, t{=}1] - E[Y \mid s{=}0, g{=}1, t{=}0] = \lambda_{1,1} - \lambda_{1,0} + \delta_{0,1} - \delta_{0,0}, \tag{8.16}$$
$$E[Y \mid s{=}1, g{=}0, t{=}1] - E[Y \mid s{=}1, g{=}0, t{=}0] = \lambda_{0,1} - \lambda_{0,0} + \delta_{1,1} - \delta_{1,0}, \tag{8.17}$$
$$E[Y \mid s{=}1, g{=}1, t{=}1] - E[Y \mid s{=}1, g{=}1, t{=}0] = \lambda_{1,1} - \lambda_{1,0} + \delta_{1,1} - \delta_{1,0} + \rho. \tag{8.18}$$

Second difference (same state, across occupations; the state by time trend cancels):

$$(8.16) - (8.15) = \lambda_{1,1} - \lambda_{1,0} - (\lambda_{0,1} - \lambda_{0,0}), \tag{8.19}$$
$$(8.18) - (8.17) = \lambda_{1,1} - \lambda_{1,0} - (\lambda_{0,1} - \lambda_{0,0}) + \rho. \tag{8.20}$$

Each second difference is itself a DID: the cashier versus teacher gap change within a state. The occupation by time trend survives, but it is the *same* in both states.

Third difference (across states):

$$(8.20) - (8.19) = \rho. \tag{8.21}$$

The occupation trend cancels, and only the policy effect remains. DDD is easily understood as a difference between two DIDs: the cashier-minus-teacher DID in New Jersey minus the cashier-minus-teacher DID in Pennsylvania.

8.8 Conclusion

When we observe the same units across time, we have panel data, and the fixed effect toolkit lets us cancel out all time invariant confounders. The FE, dummy, and FD regressions all do this; FE and dummy regressions are numerically identical, FE and FD coincide with two periods but differ beyond that, and the choice between FE and FD depends on the serial correlation structure of the errors.

The main assumption for DID is parallel trends. Traditionally we assess it in two steps: draw the picture, and run the event study. But a statistical pre-test that minimizes the type I error rate mechanically inflates the type II error rate, and Roth (2022) shows that in realistic settings the power of pre-trend tests is low, while conditioning on passing them distorts both variance and bias. Be careful using them: check the power,

do sensitivity analysis, and above all validate the assumption with economic knowledge of the setting.

When it is hard to find any control group, you can create one with the synthetic control method: two sets of weights, one over characteristics (v) and one over donor units (w), combine the donor pool into an artificial control that mimics the treated unit's pre-treatment behavior. Alternatively, if there is a further layer of difference that gives you exogeneity, such as an occupation group untouched by the policy, triple differences absorbs state specific trends with state by time fixed effects and identifies the effect from the triple interaction.

Throughout, keep asking the identification source question: which variation in the data is identifying the parameter? The answer determines the assumption you need, the interpretation you may claim, and the defense you must give. In the next chapter we stay with panel data and confront the staggered adoption case, where treatment starts at different times for different units and the TWFE regression turns out to hide serious problems.

Homework Problems

1. **FE, Dummy, and FD Estimators.**

   Consider the fixed effect model $Y_{it} = \alpha + \rho D_{it} + A_i'\gamma + \epsilon_{it}$ with $T = 2$ periods (no additional covariates).

   (a) Write down the within (FE) transformation of the model and the first difference (FD) transformation of the model.

   (b) Show that with $T = 2$ the two transformations are identical up to a constant scaling, so the FE and FD estimators of ρ coincide.

   (c) Now suppose $T > 2$. Suppose the errors ϵ_{it} are i.i.d. across t. Show that the first differenced errors $\Delta\epsilon_{it}$ are serially correlated, and compute $\text{Corr}(\Delta\epsilon_{it}, \Delta\epsilon_{i,t-1})$.

   (d) Suppose instead that ϵ_{it} follows a random walk, $\epsilon_{it} = \epsilon_{i,t-1} + u_{it}$ with u_{it} i.i.d. Which estimator, FE or FD, would you prefer now, and why?

2. **Triple Differences Identification.**

Consider the 3WFE model (8.14) with two states ($s \in \{0,1\}$), two occupation groups ($g \in \{0,1\}$), and two periods ($t \in \{0,1\}$), where $D_{it} = 1$ only if $s = 1$, $g = 1$, $t = 1$.

(a) Derive the four within-cell time differences (8.15)–(8.18) from the model, explaining why the individual effect α_i cancels in each.

(b) Form the two second differences and explain, in words, what each one holds fixed and what remains.

(c) Show that the third difference recovers ρ, and state precisely which shocks the DDD design allows for that the simple two state DID does not.

(d) Give one example of a shock that would still bias the DDD estimate.

3. **Data Splitting in Synthetic Control.**

The cross-validation procedure of the synthetic control method divides the pre-treatment sample into two parts: a training period and a validation period. Explain the reason why we split the data into these two parts. No math!

Practice Example: A DID Simulation with Event Study

In this practice example we simulate a panel with a policy that satisfies parallel trends, estimate the policy effect by TWFE DID, and then run the event study regression to verify the flat pre-trend and trace the dynamic effect.

Data generating process. Consider $J = 40$ states observed over $T = 12$ years. States $1, \ldots, 20$ are treated: a policy takes effect in year 8 and stays on. The outcome is

$$Y_{it} = \alpha_i + \lambda_t + \delta_{\tau(t)} D_i \cdot \mathbf{1}(t \geq 8) + \epsilon_{it}, \qquad \epsilon_{it} \sim \mathcal{N}(0,1),$$

where $\alpha_i \sim \mathcal{N}(0,4)$ are state effects, λ_t is a common trend, D_i indicates the treated group, and the dynamic effects at event time $\tau = t - 8$ are

$$(\delta_0, \delta_1, \delta_2, \delta_3, \delta_4) = (2.0,\ 3.0,\ 3.5,\ 4.0,\ 4.0),$$

with $\delta_\tau = 0$ for $\tau < 0$ (parallel trends holds by construction). The static DID target is

the average post-period effect, $\bar{\delta} = 3.3$.

Analysis steps. The do-file (`src/chapter 8/ch8_did.do`) performs:

1. Simulate the panel and save the dataset.
2. Descriptive step: compute and list group-by-year mean outcomes, the raw ingredients of the trends plot.
3. Static DID: regress Y on the treatment dummy with state and year fixed effects; compare $\hat{\rho}$ to $\bar{\delta} = 3.3$.
4. Manual double difference: compute the four cell means (treated/control, pre/post) and verify that the double difference matches the regression estimate.
5. Event study: regress Y on event time dummies interacted with the treated indicator, omitting $\tau = -1$; verify the pre-period coefficients are near zero and the post-period coefficients track the true $(\delta_0, \ldots, \delta_4)$. The do-file also exports the coefficient plot to `src/chapter 8/ch8_event_study.png`; it reproduces the design of Figure 8.2 with its own random draw.
6. Cluster standard errors at the state level and compare with unclustered ones. In the baseline panel the errors are i.i.d., so the two standard errors come out similar. The do-file then re-simulates the same panel with AR(1) errors within state ($\rho = 0.8$) and repeats the comparison: the unclustered standard error falls to about half of the clustered one.
7. Key lessons summary.

Key lessons.

- Under parallel trends, TWFE DID recovers the (average) policy effect; the manual double difference and the regression coefficient agree exactly in the two-by-two collapse.
- The event study shows both halves of the design's credibility: flat, insignificant pre-period coefficients, and interpretable dynamic effects afterward.
- Clustering matters: with i.i.d. errors the clustered and unclustered standard errors are similar, but once errors are serially correlated within state, the unclustered standard error becomes misleadingly small (about half of the clustered one in the AR(1) re-simulation). Cluster at the level of the treatment variation.

Chapter 11 returns to this at length.

The do-file is located at `src/chapter 8/ch8_did.do` and the simulated dataset at `data/chapter 8/ch8_did.dta`.

Chapter 9

Causal Inference with Panel Data II

9.1 Introduction

In the previous chapter we developed the panel data toolkit around a plain vanilla DID setting: some policy is implemented at time t_0 in one set of units, the treated group, but not in the other set, the untreated group. Two groups, one switch date, one comparison. The world is rarely this tidy. Minimum wages rise in different states in different years. Broadband arrives in different counties in different months. A health insurance reform rolls out province by province. When a policy can be rolled out in different places at different times, we are in the world of *staggered DID*, and the natural reflex of every applied economist is to run the same two way fixed effect regression as before and read the policy effect off the treatment coefficient.

This chapter is about why that reflex can fail, and what to do about it. The recent econometrics literature has shown that when treatment effects are heterogeneous across groups or over time, the TWFE estimator is a weighted average of group and period level treatment effects in which *some weights can be negative*. A negative weight is not a technical curiosity: it means the regression can return a negative coefficient even when the policy helped every group in every period. The first half of the chapter follows de Chaisemartin and D'Haultfœuille (2020) and develops this result step by step: the exact weight formula, a two group example in which the weights can be computed by hand, the notion of a *forbidden comparison* that explains where negative weights come from, a diagnostic for how vulnerable your estimate is, and a corrected estimator, DID_M, that never compares a switching group to an already treated one.

The second half of the chapter turns to a different generalization of DID: what if the treatment is not a binary switch but a continuous dose? The effect of the US China trade war tariffs on Chinese employment is a leading example: regions were not simply treated or untreated, they were exposed to different tariff increases. Following Callaway et al. (2021), we will see that continuous treatments force us to distinguish two kinds of causal effects, level effects and slope effects, and that the familiar parallel trend assumption identifies only the former. Identifying the latter requires a genuinely stronger assumption, strong parallel trends, and even then the TWFE regression in a staggered setting commits the same forbidden comparisons as in the binary case.

The message of the chapter is not that DID is broken. It is that the TWFE *regression* needs to be used with caution. Linear regression is a parametric tool, and when treatment effects are heterogeneous, dynamic, or continuous, the regression may aggregate the underlying effects in ways no researcher would choose deliberately. For students who want a broader map of this fast moving literature, Roth et al. (2023) provide an excellent synthesis.

9.2 Staggered DID and the TWFE Regression

Assume the policy is implemented at the level of a group g (a state, a city, a cohort), rolling out in different periods for different groups. For individual i in group g at time t, we run the same TWFE regression as in the previous chapter:

$$Y_{igt} = \gamma_g + \lambda_t + \beta D_{gt} + \epsilon_{igt}, \tag{9.1}$$

where $D_{gt} = 1$ if group g is treated at time t.

What does β estimate? In the homogeneous treatment effect case, where the policy moves every treated cell by the same amount, β is that treatment effect, and nothing in this chapter changes the conclusion of the previous one. In the heterogeneous treatment effect case, β is a weighted sum of the average treatment effects in each group and period. That sounds harmless: a weighted average of effects is still some kind of summary of the effects. But is this in general a good estimator? The answer is **no**, and the reason is that the weights are not what you think they are.

9.3 Negative Weights in the TWFE Estimator

The key reference for this section is de Chaisemartin and D'Haultfœuille (2020), abbreviated as CD (2020) below. Their main message: when treatment effects are heterogeneous, the TWFE estimator identifies a weighted average of cell level treatment effects, but some of the weights can be negative. As a consequence, the weighted average may be negative even if every single cell level effect is positive. This section unpacks why, and the derivation deserves careful attention. I strongly encourage all readers to read the original paper.

9.3.1 Setting and Assumptions

The data have three levels: individual i, group g, and period t. Let $N_{g,t}$ be the number of observations in cell (g,t) and N the total number of observations. Treatment $D_{i,g,t}$ is binary, with potential outcomes $Y_{i,g,t}(0)$ and $Y_{i,g,t}(1)$. Averaging within a (g,t) cell gives the cell level variables

$$D_{g,t} = \frac{1}{N_{g,t}} \sum_{i \in (g,t)} D_{i,g,t}, \qquad Y_{g,t}(0) = \frac{1}{N_{g,t}} \sum_{i \in (g,t)} Y_{i,g,t}(0),$$

$$Y_{g,t}(1) = \frac{1}{N_{g,t}} \sum_{i \in (g,t)} Y_{i,g,t}(1), \qquad Y_{g,t} = \frac{1}{N_{g,t}} \sum_{i \in (g,t)} Y_{i,g,t}.$$

Five assumptions set up the result.

Assumption 9.3.1 (Full Support, Assumption 1 in CD 2020)**.** For all (g,t), $N_{g,t} > 0$: every group is observed in every period.

Assumption 9.3.2 (Sharp Design, Assumption 2 in CD 2020)**.** For all (g,t) and all i in cell (g,t), $D_{i,g,t} = D_{g,t}$: treatment is identical for everyone in the same group and time cell.

Assumption 9.3.3 (Independent Groups, Assumption 3 in CD 2020)**.** The vectors $(Y_{g,t}(0), Y_{g,t}(1), D_{g,t})_{1 \le t \le T}$ are mutually independent across groups.

Assumption 9.3.3 rules out correlation across groups but allows arbitrary correlation within a group across time. It is a weaker version of the usual i.i.d. sampling assumption, adapted to grouped panel data.

Assumption 9.3.4 (Strong Exogeneity, Assumption 4 in CD 2020)**.** For all (g, t),

$$E\big(Y_{g,t}(0) - Y_{g,t-1}(0) \mid D_{g,1}, \ldots, D_{g,T}\big) = E\big(Y_{g,t}(0) - Y_{g,t-1}(0)\big).$$

Strong exogeneity says that the whole treatment path of a group carries no information about the evolution of its untreated outcome. Treatment cannot be assigned to a group *because* that group just experienced a negative or positive shock. This is the panel analogue of ruling out an Ashenfelter dip (a transitory pre treatment drop in outcomes among those selected into treatment): both the assignment of groups to treatment and the timing of that treatment must be exogenous.

One more piece of generality before the last assumption. In a staggered adoption design, once a group is treated it stays treated. de Chaisemartin and D'Haultfœuille (2020) allow a more general case in which *exit* is also possible: some group can cancel the policy after some periods, so that $D_{g,t} = 1$ and $D_{g,t+1} = 0$. Everything below covers both cases.

Assumption 9.3.5 (Common Trends, Assumption 5 in CD 2020)**.** For $t \geq 2$, $E\big(Y_{g,t}(0) - Y_{g,t-1}(0)\big)$ does not vary across g.

This is the familiar parallel trend assumption from the previous chapter, written at the group level: potential outcomes without treatment evolve identically across groups. It is important to distinguish between Assumptions 9.3.4 and 9.3.5. Assumption 9.3.4 claims the exogeneity/randomness of the treatment across both time and group; Assumption 9.3.5 states the equality of untreated outcome growth across groups.

9.3.2 What Does the TWFE Coefficient Estimate?

The natural target parameter is the ATT, the average treatment effect across all treated observations. Define the cell level average treatment effect

$$\Delta_{g,t} = \frac{1}{N_{g,t}} \sum_{i \in (g,t)} \big[Y_{i,g,t}(1) - Y_{i,g,t}(0)\big],$$

and let N_1 be the total number of treated observations. Then the ATT is an expected weighted average of the treated cells' effects,

$$\delta^{TR} = E\big[Y_{i,g,t}(1) - Y_{i,g,t}(0) \mid D_{g,t} = 1\big] = E\Big[\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1}\,\Delta_{g,t}\Big], \tag{9.2}$$

where every treated cell receives a weight proportional to its size. All weights in (9.2) are positive. This is the benchmark against which the regression will be judged.

Now let $\hat{\beta}_{fe}$ be the OLS estimator of β in regression (9.1) and $\beta_{fe} = E(\hat{\beta}_{fe})$. The characterization of β_{fe} runs through an auxiliary regression: regress the treatment indicator itself on group and period fixed effects,

$$D_{g,t} = \alpha + \gamma_g + \lambda_t + \varepsilon_{g,t}, \tag{9.3}$$

and let $\varepsilon_{g,t}$ denote the residual of this regression. (Do not confuse $\varepsilon_{g,t}$, the residual of the treatment regression at the cell level, with ϵ_{igt}, the error term of the outcome regression; the notation follows de Chaisemartin and D'Haultfœuille, 2020.)

Theorem 9.3.1 (Theorem 1 in de Chaisemartin and D'Haultfœuille, 2020)**.** *Under Assumptions 9.3.1 to 9.3.5,*

$$\beta_{fe} = E\Big[\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1}\, w_{g,t}\,\Delta_{g,t}\Big], \qquad w_{g,t} = \frac{\varepsilon_{g,t}}{\displaystyle\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1}\,\varepsilon_{g,t}}.$$

The TWFE estimator is a weighted average of cell level average treatment effects, with weights $w_{g,t}$ that sum to one but need not be positive.

Compare Theorem 9.3.1 with the ATT in (9.2). Both are weighted averages of the same objects $\Delta_{g,t}$. The ATT weights every treated cell by its share of treated observations. The regression weights every treated cell by its share *times* $w_{g,t}$, and $w_{g,t}$ is driven by the residual $\varepsilon_{g,t}$: how much the cell's treatment status deviates from what its group average and period average would predict.

How should we interpret this weight? The regression assigns more weight to treated cells (g,t) that deviate from the average treatment level of all cells in group g and of all cells at time t. If everyone else in your group, or everyone else in your period, is not treated, but you are treated, then you are assigned a large weight: you are the

surprising cell, and OLS loves surprising variation. That may seem acceptable so far. The big issue is that $\varepsilon_{g,t}$, being a regression residual, can be *negative* for a treated cell, and then the cell's treatment effect enters β_{fe} with a negative weight.

9.3.3 A Two Group, Three Period Example

To see negative weights appear in the smallest possible setting, assume $N_{g,t}/N_{g,t-1}$ does not vary across g (a balanced panel), and consider two groups of equal size observed for three periods. Group 1 gets treated at period 3; group 2 gets treated at periods 2 and 3. Figure 9.1 shows the rollout.

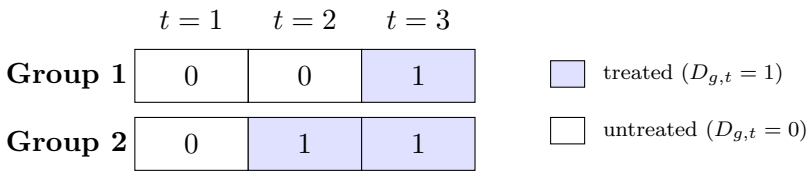


Figure 9.1: Treatment rollout in the two group, three period example. Group 2 adopts the policy at period 2, group 1 at period 3. There are three treated cells: $(1,3)$, $(2,2)$, and $(2,3)$.

With equal cell sizes, the residual of the treatment regression (9.3) has a closed form. Let

$$D_{g,\cdot} = \sum_t \frac{N_{g,t}}{N_g} D_{g,t}, \qquad D_{\cdot,t} = \sum_g \frac{N_{g,t}}{N_t} D_{g,t}, \qquad D_{\cdot,\cdot} = \sum_{(g,t)} \frac{N_{g,t}}{N} D_{g,t}$$

be the group average, period average, and grand average of treatment. Then, one can show that

$$\varepsilon_{g,t} = D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}. \tag{9.4}$$

Equation (9.4) comes from the fact that in OLS regression, group dummies lead to group means as linear predictions. In the example, group 1 is treated in one of three periods ($D_{1,\cdot} = 1/3$), group 2 in two of three ($D_{2,\cdot} = 2/3$), period 2 has one of two groups treated ($D_{\cdot,2} = 1/2$), period 3 has both ($D_{\cdot,3} = 1$), and half of all cells are

treated ($D_{\cdot,\cdot} = 1/2$). Plugging into (9.4) for the three treated cells:

$$\begin{aligned}
\varepsilon_{1,3} &= 1 - \frac{1}{3} - 1 + \frac{1}{2} = \frac{1}{6}, \\
\varepsilon_{2,2} &= 1 - \frac{2}{3} - \frac{1}{2} + \frac{1}{2} = \frac{1}{3}, \\
\varepsilon_{2,3} &= 1 - \frac{2}{3} - 1 + \frac{1}{2} = -\frac{1}{6}.
\end{aligned}$$

The residual for cell $(2,3)$ is negative. Applying Theorem 9.3.1 (each treated cell has $N_{g,t}/N_1 = 1/3$, and the denominator of the weights is $\frac{1}{3}(\frac{1}{6} + \frac{1}{3} - \frac{1}{6}) = \frac{1}{9}$), the TWFE estimand in this special case is

$$\beta_{fe} = \frac{1}{2}E[\Delta_{1,3}] + E[\Delta_{2,2}] - \frac{1}{2}E[\Delta_{2,3}]. \tag{9.5}$$

We assign a negative weight to $\Delta_{2,3}$, the treatment effect of the early adopting group in the late period. Negative weights can make results weird. Suppose all three effects are positive: $E[\Delta_{1,3}] = E[\Delta_{2,2}] = 1$ and $E[\Delta_{2,3}] = 4$. Then

$$\beta_{fe} = \frac{1}{2} \times 1 + 1 - \frac{1}{2} \times 4 = -\frac{1}{2}.$$

The policy helps every group in every period, and the regression reports a negative effect. The problem is in what the coefficient *is*, not in how precisely it is estimated.

9.3.4 Forbidden Comparisons

Where does the negative weight come from? Let us look at the same example through the lens of DID comparisons. In this design there are two switches: group 2 switches on at period 2, and group 1 switches on at period 3. Each switch invites a DID comparison of the switcher against the other group:

$$DID_1 = E(Y_{2,2}) - E(Y_{2,1}) - \big[E(Y_{1,2}) - E(Y_{1,1})\big], \tag{9.6}$$

$$DID_2 = E(Y_{1,3}) - E(Y_{1,2}) - \big[E(Y_{2,3}) - E(Y_{2,2})\big]. \tag{9.7}$$

DID_1 compares group 2's change from period 1 to 2 against group 1's change over the same window, when group 1 is still untreated. DID_2 compares group 1's change from period 2 to 3 against group 2's change over the same window, when group 2 is *already treated in both periods*. It can be proved that the TWFE estimand is the simple average

of these two comparisons:

$$\beta_{fe} = \frac{1}{2}\big(DID_1 + DID_2\big).$$

The first comparison is the familiar, clean one. The second is not, and this is the heart of the chapter. One might hope that $DID_2 = E[\Delta_{1,3}]$: the switcher's treatment effect, with group 2's change differencing out the common trend. Let us check, writing each observed outcome in potential outcome notation ($Y_{1,3} = Y_{1,3}(1)$, $Y_{1,2} = Y_{1,2}(0)$, while $Y_{2,3} = Y_{2,3}(1)$ and $Y_{2,2} = Y_{2,2}(1)$ since group 2 is treated in both periods):

$$\begin{aligned}
DID_2 &= E\big(Y_{1,3}(1)\big) - E\big(Y_{1,2}(0)\big) - \big[E\big(Y_{2,3}(1)\big) - E\big(Y_{2,2}(1)\big)\big] \\
&= \underbrace{E\big(Y_{1,3}(1)\big) - E\big(Y_{1,3}(0)\big)}_{E[\Delta_{1,3}]} + E\big(Y_{1,3}(0)\big) - E\big(Y_{1,2}(0)\big) \\
&\quad - \big[E\big(Y_{2,3}(1)\big) - E\big(Y_{2,2}(1)\big)\big] \\
&= E[\Delta_{1,3}] + \underbrace{E\big(Y_{1,3}(0)\big) - E\big(Y_{1,2}(0)\big) - \big[E\big(Y_{2,3}(0)\big) - E\big(Y_{2,2}(0)\big)\big]}_{=0 \text{ by common trends}} \\
&\quad - \big[E\big(Y_{2,3}(1)\big) - E\big(Y_{2,3}(0)\big)\big] + \big[E\big(Y_{2,2}(1)\big) - E\big(Y_{2,2}(0)\big)\big] \\
&= E[\Delta_{1,3}] - \big(E[\Delta_{2,3}] - E[\Delta_{2,2}]\big). \qquad (9.8)
\end{aligned}$$

The common trend assumption cancels the untreated outcome changes, as always. But group 2's *observed* change is a change in $Y(1)$, not in $Y(0)$, which means we cannot cancel out $E\big(Y_{1,3}(0)\big) - E\big(Y_{1,2}(0)\big)$ and $E\big(Y_{2,3}(1)\big) - E\big(Y_{2,2}(1)\big)$. Converting it into an untreated change leaves behind the difference of group 2's treatment effects across the two periods. The result is:

$$DID_2 = E[\Delta_{1,3}] - \underbrace{\big(E[\Delta_{2,3}] - E[\Delta_{2,2}]\big)}_{\text{bias term}}.$$

What does this equation mean? This DID comparison is the ATT of group 1 in period 3, *minus a bias term*: the change in group 2's treatment effect between periods 2 and 3. You are using treated cells as the control group. For this already treated group, although there is no change in treatment status from period 2 to 3, the change in its outcome is a change in $Y(1)$, and that is not comparable to a change in $Y(0)$: changes in $Y(1)$ additionally contain changes in the treatment effect across time. If group 2's effect is growing, the growth is subtracted from group 1's estimated effect, and $E[\Delta_{2,3}]$

enters β_{fe} with a negative sign. That is exactly the $-\frac{1}{2}E[\Delta_{2,3}]$ in (9.5). Figure 9.2 contrasts the two comparisons.

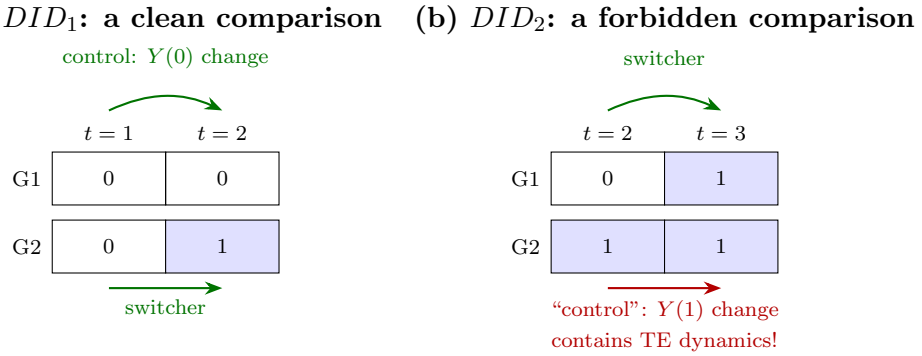


Figure 9.2: The two DID comparisons hidden inside the TWFE estimator. In panel (a), the control group is untreated in both periods, so its change identifies the common trend in $Y(0)$. In panel (b), the "control" group is treated in both periods; its observed change mixes the common trend with the change in its own treatment effect, which then enters the estimate with a negative sign.

Forbidden Comparisons

Do not use continuously treated groups as control groups. An already treated group's outcome change is a change in $Y(1)$: it contains the evolution of that group's treatment effect on top of the common trend. A DID comparison against such a group subtracts another cell's treatment effect dynamics from the switcher's effect. This is called a *forbidden comparison*, and it is the mechanism that produces negative weights in TWFE.

In general, which kinds of cells are more likely to receive negative weights? Go back to the residual formula (9.4): $\varepsilon_{g,t}$ turns negative for a treated cell when the group average $D_{g,\cdot}$ and the period average $D_{\cdot,t}$ are large. A treated cell is therefore more likely to have a negative weight if

1. it sits at a period when many groups are treated, and
2. it sits in a group that is treated for many periods.

This is the exact opposite of who gets a large positive weight: if everyone else in your group and everyone else in your year is treated, you are assigned a negative weight, just like group 2 in period 3 of the example. The intuition is that these cells are the ones most often used as "controls" in the regression's implicit comparisons: many chances

for forbidden comparisons. In a staggered adoption design, the dangerous cells are the *early adopting groups* in *late periods*, which is unfortunate, because in applications those are often exactly the cells where the policy has had time to develop its full effect.

9.3.5 What Should We Do

So, what should we do to deal with this negative weight issue? The first step is diagnostic: check how sensitive your result is to treatment effect heterogeneity. The second step is to develop a new estimator free from forbidden comparisons.

9.3.5.1 Step 1: Check How Sensitive Your Estimate Is

Sensitivity Diagnostic for TWFE

1. Compute the weights $w_{g,t}$ from Theorem 9.3.1 and see whether some of them are negative.
2. Divide $|\hat{\beta}_{fe}|$ by the standard deviation of the weights. This ratio is the minimal value of the standard deviation of the cell level ATEs across (g,t) cells under which the true ATT could have the opposite sign from your estimate.
3. If there are many negative weights, or if $|\hat{\beta}_{fe}|/sd(w)$ is small, do not use TWFE: the estimator is too vulnerable to treatment effect heterogeneity in your design.

In practice, the Stata package `twowayfeweights` computes the weights and the threshold for you. The $DIDM$ estimator and its placebo test are implemented in the Stata packages `did_multiplegt` and `fuzzydid`; the package `did_multiplegt_dyn` is the current successor of `did_multiplegt` and covers dynamic effects as well.

The logic of the ratio in step 2 is worth spelling out. If the cell level effects were homogeneous, the weights would not matter: any weights summing to one deliver the common effect. Heterogeneity is what lets negative weights do damage, and the more the effects vary across cells, the less negativity is needed to flip the sign of the weighted average. The ratio answers: how much heterogeneity, measured by the standard deviation of the cell effects, would be enough? If a modest amount suffices, your headline sign is fragile.

9.3.5.2 Step 2: The DID_M Estimator

If you have many negative weights, or the threshold value for flipping the sign is small, use a different estimator. de Chaisemartin and D'Haultfœuille (2020) construct a new estimator, called DID_M, that only compares switching cells to stable ones. The target parameter changes accordingly: instead of the ATT over all treated cells, DID_M targets the ATE of *switching* cells,

$$\delta^s = E\Big[\frac{1}{N_s}\sum_{(i,g,t):\, t\geq 2,\, D_{g,t}\neq D_{g,t-1}} \big[Y_{i,g,t}(1) - Y_{i,g,t}(0)\big]\Big], \tag{9.9}$$

where $N_s = \sum_{(g,t):\, t\geq 2,\, D_{g,t}\neq D_{g,t-1}} N_{g,t}$ is the number of observations changing treatment status between $t-1$ and t. Switching cells include *joiners* (from untreated to treated) and *leavers* (from treated to untreated).

Four additional assumptions are needed.

Assumption 9.3.6 (Strong Exogeneity for $Y(1)$, Assumption 9 in CD 2020)**.** For all (g,t), $E\big(Y_{g,t}(1) - Y_{g,t-1}(1) \mid D_{g,1},\ldots,D_{g,T}\big) = E\big(Y_{g,t}(1) - Y_{g,t-1}(1)\big)$.

Assumption 9.3.7 (Common Trends for $Y(1)$, Assumption 10 in CD 2020)**.** For $t \geq 2$, $E\big(Y_{g,t}(1) - Y_{g,t-1}(1)\big)$ does not vary across g.

Assumptions 9.3.6 and 9.3.7 mirror strong exogeneity and common trends, but for the *treated* potential outcome. Assumption 9.3.7 imposes some homogeneity on how treatment effects evolve. These two assumptions allow one to reconstruct the missing potential outcome of leavers, and they are needed *only when the design has exits*. In a staggered adoption design with no leavers, they are not necessary.

Assumption 9.3.8 (Existence of Stable Groups, Assumption 11 in CD 2020)**.** (i) If there is a group g with $D_{g,t-1} = 0$ and $D_{g,t} = 1$, there exists a group g' with $D_{g',t-1} = D_{g',t} = 0$. (ii) If there is a group g with $D_{g,t-1} = 1$ and $D_{g,t} = 0$, there exists a group g' with $D_{g',t-1} = D_{g',t} = 1$.

Assumption 9.3.9 (No Spillover, Assumption 12 in CD 2020)**.** For all (g,t), $E\big(Y_{g,t}(0) \mid \mathbf{D}\big) = E\big(Y_{g,t}(0) \mid D_g\big)$ and $E\big(Y_{g,t}(1) \mid \mathbf{D}\big) = E\big(Y_{g,t}(1) \mid D_g\big)$: a group's outcome depends only on its own treatment path, not on other groups' treatments.

Assumption 9.3.8 guarantees that every switch has a valid comparison group available: whenever someone joins, some group stays untreated; whenever someone leaves,

some group stays treated. The estimator is built from exactly these comparisons. Let $N_{d,d',t} = \sum_{g:\, D_{g,t}=d,\, D_{g,t-1}=d'} N_{g,t}$ count the observations with treatment d at t and d' at $t-1$, and define two families of DID comparisons:

$$DID_{+,t} = \sum_{g:\, D_{g,t}=1,\, D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,0,t}} \big(Y_{g,t} - Y_{g,t-1}\big) - \sum_{g:\, D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}} \big(Y_{g,t} - Y_{g,t-1}\big), \tag{9.10}$$

$$DID_{-,t} = \sum_{g:\, D_{g,t}=D_{g,t-1}=1} \frac{N_{g,t}}{N_{1,1,t}} \big(Y_{g,t} - Y_{g,t-1}\big) - \sum_{g:\, D_{g,t}=0,\, D_{g,t-1}=1} \frac{N_{g,t}}{N_{0,1,t}} \big(Y_{g,t} - Y_{g,t-1}\big). \tag{9.11}$$

$DID_{+,t}$ is the DID of joiners against stable untreated groups: the standard comparison, with a genuinely untreated control. $DID_{-,t}$ is the DID of stable treated groups against leavers: it recovers the leavers' effect by comparing their change to that of groups that remain treated, which is where Assumptions 9.3.6 and 9.3.7 come in. The estimator aggregates all switches:

$$DID_M = \sum_{t=2}^{T} \Big(\frac{N_{1,0,t}}{N_s}\, DID_{+,t} + \frac{N_{0,1,t}}{N_s}\, DID_{-,t} \Big). \tag{9.12}$$

Be careful: under additional Assumptions 9.3.6 and 9.3.7, $DID_{-,t}$ will not cause a forbidden comparison even if we seem to use already treated cells, because in forbidden comparisons we compare a newly treated group whose counterfactual state is untreated, to an incorrect "control" group whose state is being continuously treated. However, here for leavers, we compare a leaver whose counterfactual state is being continuously treated, to a correct control group whose state is also being continuously treated. The key is to have the counterfactual of the targeted group match the control group.

Theorem 9.3.2 (Theorem 3 in de Chaisemartin and D'Haultfœuille, 2020)**.** *Under Assumptions 9.3.1, 9.3.2, 9.3.4, 9.3.5, and 9.3.6 to 9.3.9,*

$$E[DID_M] = \delta^s.$$

The DID_M estimator is a weighted average of joiners' and leavers' treatment effects, and it is an unbiased estimator of δ^s, the ATE of all switching cells.

Three remarks on the properties of DID_M. First, it is consistent and asymptotically

normal, so inference proceeds as usual. Second, it is non-parametric: it uses only cell means and never fits a global linear model, and it is therefore less efficient than TWFE when TWFE is valid. This is the bias and variance tradeoff from Chapter 3 again, now between estimators rather than tuning parameters: TWFE buys precision with a parametric structure that fails under heterogeneity, while DID_M pays variance for robustness. Third, a placebo test of the common trend assumption comes almost for free. The basic idea is to compare the outcome's evolution from $t-2$ to $t-1$ between the groups that change and do not change their treatment status from $t-1$ to t: compute $DID_{+,t}$ and $DID_{-,t}$ after replacing $Y_{g,t}$ with $Y_{g,t-1}$ and $Y_{g,t-1}$ with $Y_{g,t-2}$. If common trends holds, the switch at time t should have no echo in the earlier window, so you should expect zeros in this placebo. This is the pre-trend logic of the previous chapter adapted to staggered designs, and the warnings of Roth (2022) about pre-test power apply here as well.

The basic idea of this DID_M estimator is to open the black box of the traditional TWFE estimator. Rather than giving it a functional form (linearity) restriction, the DID_M estimator starts from the non-parametric definition of the target treatment effect and manually excludes all forbidden comparisons, which are incorrectly contained in the TWFE estimator. There are many other new estimators proposed by econometricians to deal with the staggered DID issue, such as Callaway and Sant'Anna (2021); Borusyak et al. (2024); Sun and Abraham (2021). They are partly different for various DID settings, but all share a similar idea about estimator construction as in DID_M. We do not have space to introduce them one by one, so for interested readers, please refer to the original papers for details.

9.3.6 Takeaways

Main Takeaways of de Chaisemartin and D'Haultfœuille (2020)

- In general, TWFE is not a good estimator in settings with heterogeneous treatment effects: it may assign negative weights to some group and period ATEs.
- If your design has periods when many groups are treated, or groups treated for many periods, be careful: those cells are the negative weight candidates.
- In practice: (i) calculate the weights and the threshold value for flipping the sign; (ii) if there are many negative weights or the threshold is small,

use DID_M (Stata: `did_multiplegt`) or other new estimators rather than TWFE; (iii) if you use new estimators, implement the new placebo test to validate the parallel trend assumption.
- The argument is not specific to binary treatments: it applies to any ordered treatment D.

Roth et al. (2023) give a very clear and comprehensive literature review on the new development of DID method these days. They also prepare a table with different new estimators and their corresponding Stata packages, which would be extremely useful for empirical research.

9.4 DID with a Continuous Treatment

So far the treatment was a switch: on or off. Many treatments of first order interest are instead *doses*. What is the effect of the US China trade war tariff on China's employment? Regions differ in how much tariff exposure they received, not merely in whether they received any. Import competition, minimum wage levels, pollution regulations, transfer amounts: continuous treatments are everywhere.

In general, we will see that TWFE with a continuous treatment is much more complicated than the binary case. This is not surprising as we have shown the growing complexity for multivalued/continuous IV in Chapter 6. In TWFE settings, we also need additional assumptions to identify meaningful causal effects, and the staggered version suffers from the same forbidden comparison problem we just diagnosed. More fundamentally, the continuous case forces us to think about what "the effect" even means in the potential outcome framework, and the answer is usually complicated. The key reference is Callaway et al. (2021). They define two types of causal effects, a level effect (dose d versus 0) and a slope effect (dose d versus d'), work out the assumptions needed to identify each non-parametrically in a simple two period case, analyze what the TWFE estimator delivers, and extend the results to multiple periods and staggered designs.

9.4.1 Setting: Two Periods, One Dose

Consider first the cleanest case: two periods and a one time treatment. We have periods $t-1$ and t; units receive a treatment dose D_i in period t and nothing in period $t-1$. The potential outcome of unit i at time s under dose d is $Y_{is}(d)$.

Assumption 9.4.1 (Random Sampling, Assumption 1 in Callaway et al., 2021)**.** We observe i.i.d. samples.

Assumption 9.4.2 (Continuous Treatment, Assumption 2-Cont in Callaway et al., 2021)**.** The support of the dose is $\mathcal{D} = \{0\} \cup \mathcal{D}_+$ with $\mathcal{D}_+ = [d_L, d_U]$, $0 < d_L < d_U < \infty$, $P(D = 0) > 0$, and $f_D(d) > 0$ for all $d \in \mathcal{D}_+$.

Assumption 9.4.3 (No Anticipation, Assumption 3 in Callaway et al., 2021)**.** $Y_{it-1} = Y_{it-1}(0)$ and $Y_{it} = Y_{it}(D_i)$.

There is a group of genuinely untreated units (dose zero) and a continuum of treated doses between a lowest dose d_L and a highest dose d_U. The untreated group will play the role Pennsylvania played in the previous chapter; the new difficulty lives entirely inside the treated continuum.

9.4.2 Level Effects and Slope Effects

With a binary treatment there is one comparison: treated versus untreated. With a continuous treatment, the definition of a causal effect becomes much richer, because you can compare dose d not only with 0 but also with a nearby dose d'. Callaway et al. (2021) organize the possibilities into two families.

Definition 9.4.1 (Level Effects)**.** The causal effect of dose d relative to no treatment is $Y_t(d) - Y_t(0)$. Its averages are

$$ATT(a \mid b) = E\big[Y_t(a) - Y_t(0) \mid D = b\big], \qquad ATE(d) = E\big[Y_t(d) - Y_t(0)\big].$$

$ATT(a \mid b)$ is the average effect of *potential* dose a for the units that *actually* experienced dose b; $ATE(d)$ is the average effect of dose d over the whole population.

Definition 9.4.2 (Slope Effects: Average Causal Response)**.** The causal response to a marginal increase in the dose is the derivative of the potential outcome function. Its

averages are

$$ACRT(d \mid d) = \frac{\partial\, E[Y_t(l) \mid D = d]}{\partial l}\Big|_{l=d}, \qquad ACR(d) = \frac{\partial\, E[Y_t(d)]}{\partial d}.$$

$ACRT(d \mid d)$ is the average causal response for the group that actually experienced dose d: what is the impact for people who got dose d of getting a little bit more? $ACR(d)$ is the same object averaged over everyone. The pair is analogous to ATT and ATE.

Applied papers routinely target slope effects, because the policy question is usually "what happens if we increase the dose," not "what happens relative to abolishing the program entirely." However, usually regression coefficients represent a weighted average of these two effects under strong assumptions. Keep the two families separate in your mind; their identification requirements turn out to be very different.

9.4.3 What Does Parallel Trends Identify?

We start with the natural generalization of the parallel trend assumption and examine what it could identify non-parametrically.

Assumption 9.4.4 (Parallel Trends, in Callaway et al., 2021)**.** For all d, $E\big[Y_t(0) - Y_{t-1}(0) \mid D = d\big] = E\big[Y_t(0) - Y_{t-1}(0) \mid D = 0\big]$.

The path of *untreated* potential outcomes would have been the same for the untreated group and for the treated group at any dose level. This is exactly the assumption of the previous chapter, restated dose by dose. Under it, level effects are identified by the familiar double difference, where $\Delta Y_t = Y_t - Y_{t-1}$.

Theorem 9.4.1 (Theorem 1 in Callaway et al., 2021)**.** *Under Assumptions 9.4.1 to 9.4.4, $ATT(d \mid d)$ is identified for all $d \in \mathcal{D}$:*

$$ATT(d \mid d) = E\big[\Delta Y_t \mid D = d\big] - E\big[\Delta Y_t \mid D = 0\big].$$

We can non-parametrically identify the level effect under the parallel trend assumption, in a DID fashion: compare the outcome change of the dose d group with the outcome change of the untreated group. So far, the continuous case looks like a family of binary DIDs, one per dose. The trouble starts with slope effects.

Proposition 9.4.1 (Proposition 3(a) in Callaway et al., 2021)**.** *Under Assumptions 9.4.1 to 9.4.4, in general* $ACRT(d \mid d)$ *is* not *identified:*

$$\frac{\partial\, E[\Delta Y_t \mid D = d]}{\partial d} = ACRT(d \mid d) + \underbrace{\frac{\partial\, ATT(d \mid l)}{\partial l}\Big|_{l=d}}_{\text{selection bias}}.$$

The local comparison of outcome paths across nearby doses mixes the causal response with a selection bias term: the marginal change in the treatment effect of dose d across the groups that received different doses. Why does parallel trends not help here? Because for $ACRT(d \mid d)$ you are comparing dose d with a nearby dose l, not with zero. Parallel trends disciplines the *untreated* potential outcomes of the two groups, $Y_t(0) - Y_{t-1}(0)$, and says nothing about whether the dose d group and the dose l group would respond identically *if both were given dose d*. To see the bias term as a difference, note that for d' close to d the bias term $\partial\, ATT(d \mid l)/\partial l\,\big|_{l=d}$ is proportional to the difference

$$E\big[Y_t(d) - Y_{t-1}(0) \mid D = d\big] - E\big[Y_t(d) - Y_{t-1}(0) \mid D = d'\big],$$

with the proportionality factor $1/(d - d')$ coming from the definition of a derivative: when we subtract the observed path of the dose d' group from that of the dose d group, we get both the causal response and the difference in how the two groups would have responded to the *same* dose. Here we have $Y_t(d)$ but not $Y_t(0)$, which cannot be canceled out by parallel trend assumption.

Take the effect of U.S. tariff exposure on the employment rate in different Chinese provinces as an example. Assume that Shanghai is hit by 1%, and Beijing by 0.99%. When we want to estimate ACRT for Shanghai, we subtract employment rate growth in Beijing from that in Shanghai, and divide it by $1\% - 0.99\%$. This is the sample analogue of $\frac{\partial\, E[\Delta Y_t|D=d]}{\partial d}$. The assumption we really need is that Shanghai and Beijing would have grown at the same rate if they were both hit by 1% as Shanghai does in the real world. However, the parallel trend assumption only says that Shanghai and Beijing would have grown at the same rate if they were *not* hit by the U.S. tariff. The discrepancy between these two arguments leads to our selection bias, that is, Beijing and Shanghai may have different policy impact even if they were both hit by 1% tariff. This bias can be large if the Roy selection mechanism (see Chapter 6) is at play.

Therefore, to identify ACRT or ACR, we need some additional exogeneity or homogeneity in how doses were assigned. That is what the next assumption imposes.

Assumption 9.4.5 (Strong Parallel Trends, Assumption 5 in Callaway et al., 2021)**.** For all d',

$$E\big[Y_t(d') - Y_{t-1}(0)\big] = E\big[Y_t(d') - Y_{t-1}(0) \mid D = d'\big].$$

Strong parallel trends says: for every dose d', the average change in outcomes that the *whole population* would experience if assigned dose d' equals the average change actually experienced by the group that received dose d'. It imposes homogeneity of treatment effect paths across dose groups, which ordinary parallel trends never did. Note the exact strength of the statement: it is not as strong as demanding parallel trends between group $D = d'$ and *every* other group at dose d'; it demands parallel trends between group $D = d'$ and the *average of all* groups at dose d'. Under it, the slope effect for everyone becomes identified. But here, we identify ACR, the average of all units, but not ACRT.

Proposition 9.4.2 (Proposition 3(b) in Callaway et al., 2021)**.** *Under Assumptions 9.4.1 to 9.4.3 and 9.4.5, $ACR(d)$ is identified:*

$$\frac{\partial\, E[\Delta Y_t \mid D = d]}{\partial d} = ACR(d).$$

Strong Parallel Trends Is Not Testable by Pre-Trends

The strong parallel trend assumption restricts *treated* potential outcomes: how different dose groups would have responded to the same dose. Pre-treatment periods contain no doses at all, so no pre-trend plot or placebo regression can ever speak to it. You need not only quasi random assignment of *whether* a unit is treated, but quasi random assignment of *how much*. Defending this assumption is an economic argument about how doses were assigned, not a statistical test.

9.4.4 The TWFE Estimator with a Continuous Treatment

Now we consider the causal interpretation of the traditional TWFE estimator, which in the two period case is simply the OLS regression of ΔY_t on the dose. Callaway et al. (2021) show that its estimand is an integral over doses.

Theorem 9.4.2 (Theorem 3(a) in Callaway et al., 2021)**.** *Under Assumptions 9.4.1 to 9.4.4,*

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1(l) \left[ACRT(l \mid l) + \underbrace{\frac{\partial\, ATT(l \mid h)}{\partial h}\bigg|_{h=l}}_{\text{selection bias}} \right] dl + w_0 \frac{ATT(d_L \mid d_L)}{d_L},$$

where (i) $w_1(l) \geq 0$ *and* $w_0 > 0$*, and (ii)* $\int_{d_L}^{d_U} w_1(l)\, dl + w_0 = 1$*.*

Read the three ingredients. The first term inside the integral is the average causal response for the treated, accumulated across the dose distribution from d_L to d_U. The third term is the level effect of having the lowest dose, d_L versus 0, scaled by the dose. And the second term is the selection bias from Proposition 9.4.1, which contaminates the estimand whenever strong parallel trends fails. Under ordinary parallel trends alone, the TWFE coefficient is a mixture of causal responses *and* selection biases, with positive weights. Imposing strong parallel trends cleans out the bias.

Theorem 9.4.3 (Theorem 3(b) in Callaway et al., 2021)**.** *Under Assumptions 9.4.1 to 9.4.3 and 9.4.5,*

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1(l)\, ACR(l)\, dl + w_0 \frac{ATE(d_L)}{d_L}.$$

Under strong parallel trends, the selection bias is eliminated, and TWFE delivers a positively weighted average of average causal responses plus a lowest dose level term. Even this best case deserves two caveats. The weights $w_1(l)$ are driven by the dose distribution, not by policy relevance, so the regression summarizes the dose response curve in a way you did not choose. And remember, the result is for the clean two period design.

What if we extend to multiple periods and staggered adoption of doses? Under strong parallel trends, Callaway et al. (2021) show that the TWFE coefficient is composed of four kinds of comparisons:

1. paths of outcomes for units treated at the same time but with different doses;
2. paths of outcomes in early treated groups relative to later treated groups, in periods before the later group is treated;
3. paths of outcomes between later treated groups and *already treated* groups, in

the post treatment period of the later group;

4. paths of outcomes between early treated and later treated groups in their common post treatment periods, relative to their common pre periods.

The first two are fine. Comparisons (iii) and (iv) are forbidden comparisons: the "control" group's change is not $Y_t(0) - Y_{t-1}(0)$ but a change in treated outcomes, exactly the disease diagnosed in the binary staggered case. The same issue as in de Chaisemartin and D'Haultfœuille (2020) reappears, now with a continuous dose on top.

9.4.5 What Should We Do?

The overall conclusion of Callaway et al. (2021) is mixed. In general, it is hard to identify meaningful causal effects in a DID fashion with a continuous treatment. Causal level effects on the treated are non-parametrically identified under the common parallel trend assumption, but causal slope effects are identified only under strong parallel trends, an assumption about treated potential outcomes that no pre-trend test can probe. You need not only random assignment of treatment, but random assignment of the dose amount. In the two period case, TWFE delivers a weighted average of causal responses only under strong parallel trends; in the staggered case, TWFE suffers from negative weights and forbidden comparisons *even* under strong parallel trends. Callaway et al. (2021) provide a new non-parametric estimator to avoid the negative weight issue. This can be used as an important robustness check. For interested readers, please refer to the original paper.

Practical Guidance for Continuous DID

- Be explicit about which estimand you target: a level effect (d versus 0) or a slope effect (d versus d'). They require different assumptions.
- For slope effects, state the strong parallel trend assumption openly and defend the dose assignment economically. Remember that no pre-trend test can validate it.
- Consider estimating the dose response curve non-parametrically using the new estimator proposed by Callaway et al. (2021) rather than reading a single coefficient off a TWFE regression. Callaway et al. (2021) do not provide a Stata package, so this requires some hand work.
- Alternatively, use a structural or theoretical model to interpret the estimates.

When the design cannot deliver assumption free identification, theory is what disciplines the estimation and interpretation.

9.5 Conclusion

This chapter delivered two warnings about the workhorse TWFE regression, and one broader lesson.

The first warning concerns staggered adoption with a binary treatment. When treatment effects are heterogeneous, the TWFE coefficient is a weighted average of cell level effects in which early adopters in late periods can receive negative weights, because the regression implicitly uses already treated groups as controls. Those forbidden comparisons subtract treatment effect dynamics from the estimate, and can flip its sign outright. The remedy is to diagnose first, compute the weights and the sign flipping threshold, and to switch to new non-parametric estimators proposed by recent papers when the diagnosis is bad, validating common trends with its placebo test.

The second warning concerns continuous treatments. Level effects behave like a family of binary DIDs and are identified under ordinary parallel trends. Slope effects, the objects most policy questions actually ask about, are identified only under strong parallel trends, which restricts treated potential outcomes and cannot be tested with pre-trends. And in staggered dose designs, the forbidden comparison problem returns on top of everything else.

The broader lesson closes the arc that began in Chapter 2. Linear regression is, after all, a parametric method that imposes strong functional form assumptions. It is a simple and elegant statistical tool, and for clean designs it is exactly the right one. But when things become more and more complicated, heterogeneous, dynamic, continuous, regression may not be capable of capturing the data patterns, and it can give weird results while looking perfectly healthy. There are two fundamental ways out. The first is non-parametric tools, which give you enough flexibility to capture complicated patterns; that is the toolkit of Chapters 2 and 3, and it is exactly what DID_M and the non-parametric dose response methods do. The second is structural modeling, which uses economic theory to regulate and rationalize the data; we will meet it when we study discrete choice models in Chapters 12 and 13.

Homework Problems

1. **The Clean Comparison.**

   In the two group, three period example of Section 9.3.3, the TWFE estimand is the average $\beta_{fe} = \frac{1}{2}(DID_1 + DID_2)$, and we derived in the text that $DID_2 = E[\Delta_{1,3}] - (E[\Delta_{2,3}] - E[\Delta_{2,2}])$.

   Show what DID_1 is: starting from its definition (9.6), express it in terms of the $E[\Delta_{g,t}]$, using the common trend assumption. Explain what the difference is between the derivation of DID_1 and that of DID_2: why can some terms be canceled out in the derivation of DID_1 but not in the derivation of DID_2? This exercise should give you a deeper understanding of why using treated cells as controls is dangerous.

2. **Plain Vanilla DID.**

   Will a plain vanilla DID design, like Card and Krueger (1994) with one treated state, one control state, and one adoption date, suffer from the same negative weight issue when we use the TWFE estimator? Explain your answer in words. Do not use math.

3. **Computing the Weights in a Two Group, Four Period Design.**

   Two equal sized groups are observed for four periods with balanced cells. Group 1 is treated only in period 4; group 2 is treated in periods 3 and 4.

   (a) Compute $D_{g,.}$, $D_{.,t}$, and $D_{.,.}$, and use (9.4) to compute the residuals $\varepsilon_{g,t}$ for the three treated cells $(1,4)$, $(2,3)$, $(2,4)$.

   (b) Compute the weights $w_{g,t}$ of Theorem 9.3.1 and show that the TWFE estimand is

   $$\beta_{fe} = \tfrac{1}{3}E[\Delta_{1,4}] + E[\Delta_{2,3}] - \tfrac{1}{3}E[\Delta_{2,4}].$$

   Verify that the weights sum to one.

   (c) Find strictly positive values of $E[\Delta_{1,4}]$, $E[\Delta_{2,3}]$, $E[\Delta_{2,4}]$ such that $\beta_{fe} < 0$.

   (d) Which cell receives the negative weight, and how does this fit the general intuition about which cells are dangerous (groups treated for many periods, periods with many treated groups)?

Practice Example: Negative Weights and DID_M in a Staggered Simulation

In this practice example we bring the chapter's central calculation to data. We simulate the two group, three period design of Section 9.3.3 at the individual level, watch the TWFE regression return a *negative* estimate when every true effect is positive, verify the CD weight formula by hand, and then compute the DID_M estimator in an extended design where it is feasible.

Data generating process, scenario A. Two groups are observed for 3 periods, with 500 observations per group and period cell. Group 1 is treated in period 3; group 2 is treated in periods 2 and 3. The outcome is

$$Y_{igt} = \gamma_g + \lambda_t + \Delta_{g,t} D_{gt} + \epsilon_{igt}, \qquad \epsilon_{igt} \sim \mathcal{N}(0,1),$$

with group effects $(\gamma_1, \gamma_2) = (0, 2)$, a common trend $\lambda_t = 0.5\,t$, and cell treatment effects

$$\Delta_{1,3} = 1, \qquad \Delta_{2,2} = 1, \qquad \Delta_{2,3} = 4.$$

All treatment effects are positive, and common trends holds by construction. By (9.5) the TWFE estimand is $\frac{1}{2}(1) + 1 - \frac{1}{2}(4) = -0.5$.

Data generating process, scenario B. Identical, except a third, never treated group with 500 observations per period ($\gamma_3 = -1$) is added. The never treated group makes the stable group assumption hold at every switch date, so DID_M becomes computable at both switches. The switching cells are group 2 at period 2 and group 1 at period 3, both with effect 1, so $\delta^s = 1$.

Analysis steps. The do-file (`src/chapter 9/ch9_staggered_did.do`) performs:

1. Simulate scenario A and save the dataset.
2. Run the TWFE regression (9.1); verify $\hat{\beta}_{fe} \approx -0.5$ despite all cell effects being positive.
3. Verify Theorem 9.3.1 by hand: regress D_{gt} on group and period dummies at the cell level, recover the residuals $(\frac{1}{6}, \frac{1}{3}, -\frac{1}{6})$, form the weights, and reconstruct the TWFE estimate as $\frac{1}{2}\hat{\Delta}_{1,3} + \hat{\Delta}_{2,2} - \frac{1}{2}\hat{\Delta}_{2,3}$ from estimated cell effects.
4. Decompose the estimate into the two DID comparisons: compute DID_1 and

DID_2 from the six cell means, verify $\hat{\beta}_{fe} = \frac{1}{2}(DID_1 + DID_2)$, and verify the forbidden comparison formula $DID_2 \approx \Delta_{1,3} - (\Delta_{2,3} - \Delta_{2,2})$.

5. Simulate scenario B; compute $DID_{+,2}$ (group 2 against the stable untreated groups 1 and 3) and $DID_{+,3}$ (group 1 against the stable untreated group 3) and aggregate to $DID_M \approx 1 = \delta^s$. Note why scenario A cannot deliver DID_M at the second switch: with only two groups, no stable untreated group exists at $t = 3$ (Assumption 9.3.8 fails there).
6. Run the DID_M placebo comparison in scenario B as a common trend check.
7. Key lessons summary.

Key lessons.

- A correctly specified TWFE regression on a valid DID design can return a significantly negative estimate when every true effect is positive: negative weights are not a small sample quirk but a property of the estimand.
- The CD weights are computable from the treatment layout alone, before ever touching the outcome: the diagnosis costs nothing.
- DID_M repairs the problem by only comparing switchers to stable groups, at the price of estimating a different, honestly labeled estimand, the switchers' ATE.

The do-file is located at `src/chapter 9/ch9_staggered_did.do` and the simulated datasets at `data/chapter 9/`.

Chapter 10

Regression Discontinuity Design

10.1 Introduction

Suppose we want to examine the education quality of Peking University (PKU) and Fudan University (FDU). We collect wage data for their graduates and find that the average wage for PKU graduates is 200,000 RMB per year, while the average wage for FDU graduates is 150,000 RMB per year. Does this mean that studying at PKU results in higher human capital than studying at FDU?

The answer is no, and by now the reason should be second nature: better students select into PKU. The wage gap mixes two things, the value added of the school and the initial quality of the students it admits. Self selection is always a problem in economic research. Whenever we ask whether school A is more efficient than school B, we must confront the possibility that school A simply admits students with better initial quality. How do we deal with this issue?

One route is to construct a selection model structurally. But there is another, design-based approach: the *Regression Discontinuity Design* (RDD). The intuition for RDD is simple. Admission to PKU is determined by the college entrance exam score and a sharp admission line. Consider PKU students who scored *just above* the PKU admission line, and FDU students who scored *just below* it. These two sets of students enrolled in PKU or FDU essentially by chance: a lucky guess on one multiple choice question separates them. They are therefore similar in ability, and comparing their later outcomes isolates the effect of attending PKU.

This chapter develops the RDD toolkit in four steps. Section 10.2 introduces the sharp design, where treatment is a deterministic function of a running variable, together with the practical questions of how to model the outcome away from the cutoff. Section 10.3 introduces the fuzzy design, where crossing the cutoff shifts only the *probability* of treatment, and shows that fuzzy RD is naturally implemented as an IV estimator. Section 10.4 steps back and asks the deeper questions: what causal effect does RD identify, and under exactly what assumptions? The answers come from the classic paper of Hahn et al. (2001). Section 10.5 presents an extension, the Regression Kink Design (RKD), where identification comes from a change in slope rather than a jump in level. Section 10.6 presents an RDD application, He et al. (2020), which uses a spatial RD built into China's water quality monitoring system.

10.2 Sharp RD

10.2.1 The Treatment Rule

Let us first consider the simple case: *Sharp RD*. In a sharp design, the treatment rule is deterministic. You are definitely treated if you surpass the threshold, and definitely not treated otherwise. There is no uncertainty in treatment assignment.

Formally, suppose that treatment D_i is determined by some variable x_i:

$$D_i = \mathbf{1}(x_i \geq x_0) = \begin{cases} 1, & \text{if } x_i \geq x_0, \\ 0, & \text{if } x_i < x_0. \end{cases} \tag{10.1}$$

Three pieces of terminology:

- x_i is called the *running variable* (the exam score, the vote share, the birth date);
- x_0 is a known threshold or *cutoff* (the admission line);
- D_i is a deterministic function of x_i.

The estimation idea is to compare samples just above x_0 and just below x_0. This is best understood as a special case of matching. In conventional matching, we compare samples with identical covariates: for each treated unit we find untreated units that look the same on observables. In RD, we compare samples within a small neighborhood of the treatment threshold: units just above and just below the cutoff are "matched"

on the running variable itself, and, if nothing else changes discontinuously at the cutoff, on everything that matters.

10.2.2 A Regression Model

We can write a simple model for this RD:

$$Y_i = f_0(x_i)\,\mathbf{1}(x_i < x_0) + f_1(x_i)\,\mathbf{1}(x_i \geq x_0) + \rho D_i + \epsilon_i. \tag{10.2}$$

Here $f_0(x_i)$ is the smoothing function below the threshold and $f_1(x_i)$ is the smoothing function above the threshold. Their job is to fit the trend of the outcome in the running variable *away* from the cutoff. D_i is the treatment indicator, jumping from 0 to 1 exactly at $x_i = x_0$. The coefficient ρ is the treatment effect: the part of the change in Y at the cutoff that cannot be explained by the smooth trend.

The simplest choices for the smoothing functions are linear and quadratic functions. Figure 10.1 shows the model at work, with two simulated examples. In Panel A, the conditional mean of the outcome is linear on each side of the cutoff, so linear f_0 and f_1 are the right choice: the two fitted lines recover the treatment jump at $x_0 = 0.5$ cleanly. In Panel B, the data are generated from model (10.2) with strongly non-linear branches: below the cutoff the outcome is almost flat at first and then rises steeply, a convex shape, while above the cutoff it climbs to an interior peak and then turns down, a concave shape. Fitting a quadratic smoothing function separately on each side bends with the data, and the gap between the two fitted branches at the cutoff again recovers the true jump ρ.

Panel B also carries a warning. Notice what a straight line would do there: it cannot bend with the data, so part of the curvature would be loaded into the estimated discontinuity, and $\hat{\rho}$ would be biased. The treatment effect and the smoothing functions are identified together: $\hat{\rho}$ is only as credible as the fit of f_0 and f_1 near the cutoff, and the functional form of the smoothing function is not an innocent choice. We return to this choice in the next subsection.

Remark 10.2.1 (Normalization at the cutoff). For ρ to capture the whole jump at the cutoff, the two smooth branches must be anchored to agree at x_0, that is, $f_0(x_0) = f_1(x_0)$. In practice this is imposed automatically by the standard estimating equation:

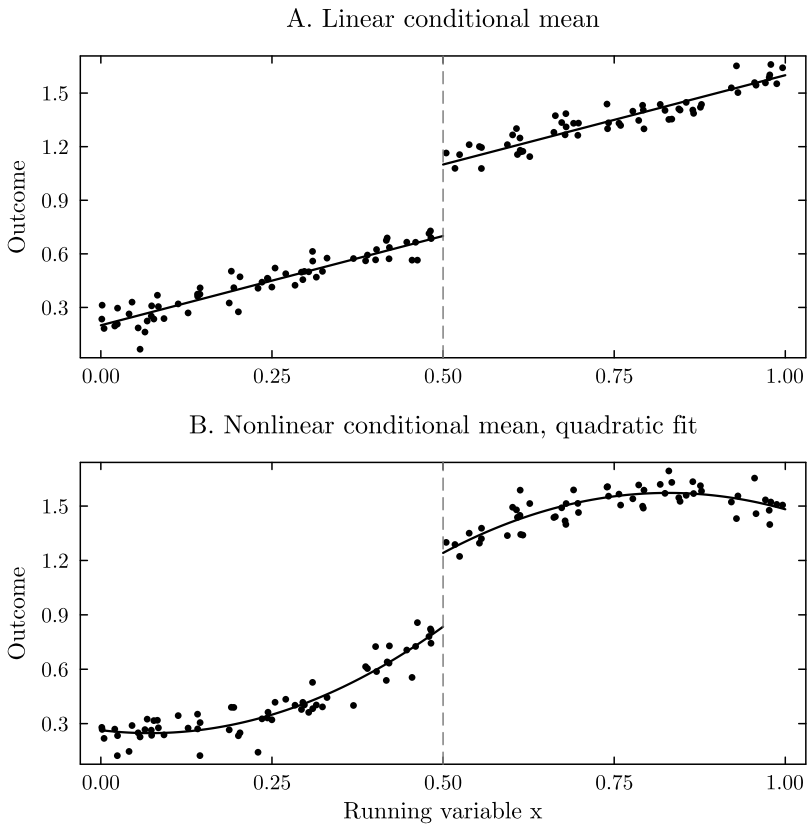


Figure 10.1: The sharp RD model (10.2) with linear and quadratic smoothing functions, following the example of Angrist and Pischke (2009, p. 255). In Panel A the conditional mean is linear on each side and the linear fits recover the true jump at $x_0 = 0.5$. In Panel B the conditional mean is strongly non-linear on each side, convex below the cutoff and concave above it, with a true jump at the threshold; the solid curves are quadratic fits estimated separately on each side, which track the curvature and recover the jump.

center the running variable at the cutoff, $\tilde{x}_i = x_i - x_0$, and run

$$Y_i = \alpha + \beta_0 \tilde{x}_i + D_i (\rho + \beta_1 \tilde{x}_i) + \epsilon_i,$$

possibly with higher order terms in $\tilde{x}_i$ on both sides. At $\tilde{x}_i = 0$ the two branches share the intercept α, and ρ is exactly the jump.

10.2.3 Choosing the Smoothing Functions

Beyond simple linear and quadratic functions, we can also use the non-parametric and semi-parametric methods introduced in Chapter 2, which are more flexible. The most recommended and commonly used one is the **Local Linear or Local Quadratic**

Regression: fit a linear or quadratic function using only observations within a bandwidth h of the cutoff, separately on each side, and read the treatment effect off the gap between the two fitted values at x_0.

As we discussed in Chapter 2, there is a bias variance tradeoff behind this choice. If you choose a complicated smoothing function, you may lose accuracy: many parameters spread over limited data produce noisy estimates. If you choose too simple a smoothing function, you may get bias: fitting straight lines to the data of Panel B in Figure 10.1 would mix the curvature into the estimated jump.

But remember one asymmetry that is specific to RD: the effective sample size is usually limited. Even if the full dataset is large, identification comes from observations in a small neighborhood around the cutoff, and only those observations carry real information about the jump. This pushes strongly in one direction: *do not use too complicated smoothing models to avoid overfitting.*

Do Not Use High Order Global Polynomials (Gelman and Imbens, 2019)

Gelman and Imbens (2019) argue that estimating RD with high order global polynomials (above third order) should be avoided, for three reasons:

- It leads to noisy estimates. High order polynomial fits assign huge and erratic effective weights to observations far from the cutoff (Runge's phenomenon): the fitted value at the boundary is driven by data noise that has nothing to do with the discontinuity.
- The RD estimate becomes very sensitive to the degree of the polynomial. Moving from a fourth to a fifth order fit can flip the sign of the estimate, with no principled way to choose between them.
- The coverage of the resulting confidence intervals is smaller than nominal: a "95 percent" interval misses the truth far more than 5 percent of the time.

The recommended choice is local linear or local quadratic regression within a bandwidth around the cutoff.

10.2.4 An Example: Lee (2008) on Incumbency Advantage

An interesting example of sharp RD is Lee (2008). The question: what is the advantage of party incumbency for reelection to the US House of Representatives? If the Democratic party holds a seat, does holding it *by itself* make the party more likely to

win the seat in the next election?

This is hard to identify, because a party may have a larger group of supporters in a district for many reasons other than incumbency. Different parties are advantaged in different regions due to ideology, history, religion, and many other factors: think of reliably blue states, reliably red states, and swing states in US presidential elections. Districts where Democrats win are simply different from districts where they lose, so a raw comparison of "held seats" and "not held seats" is contaminated by selection, exactly like the PKU and FDU comparison.

But for elections with very close results, winners and losers are similar. Whether the Democratic candidate wins a district by 0.1 percent or loses it by 0.1 percent is essentially a coin flip, and the two kinds of districts should be comparable in every other respect. Lee (2008) therefore considers the probability of a Democratic win in election t in districts where the Democratic candidate won or lost by small vote share margins in the previous election $t-1$. The running variable is the Democratic vote share margin of victory at $t-1$, and the cutoff is zero: the design is sharp because winning at $t-1$ is a deterministic function of the vote margin.

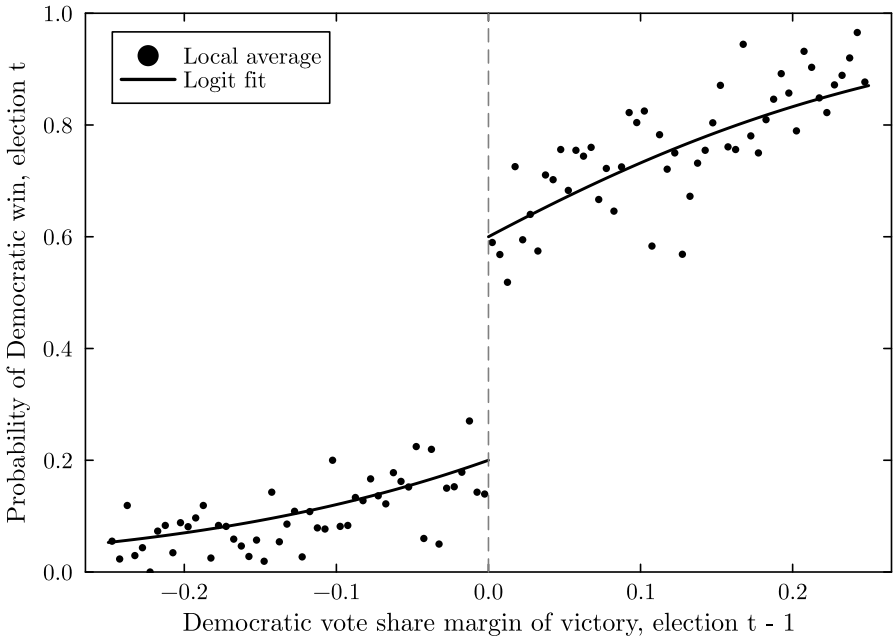


Figure 10.2: Sharp RD in Lee (2008), replicated on simulated data. Dots are local averages of the Democratic winning probability in election t, computed in narrow bins of the Democratic vote share margin at election $t-1$; solid curves are the logistic conditional probabilities on each side of the cutoff. The jump at zero is the incumbency effect.

Figure 10.2 shows the pattern. Away from zero the winning probability rises smoothly with the previous margin, reflecting persistent district preferences. At zero there is a large jump: barely winning a seat at $t-1$ raises the probability of winning at t by a large amount relative to barely losing it. That jump cannot be explained by district characteristics, which are effectively identical on the two sides of a razor thin margin. It is the causal effect of incumbency itself.

10.3 Fuzzy RD

10.3.1 Discontinuity in the Treatment Probability

A more complicated case is *Fuzzy RD*. In a fuzzy design, treatment assignment is no longer deterministic: there is uncertainty in being treated or not, and passing the threshold gives you a *larger probability* of getting treated. Perhaps scoring above the line makes admission likely but not certain; perhaps some students below the line are admitted through a special channel.

Formally, there is a discontinuity in the treatment probability, but not in treatment itself:

$$P(D_i = 1 \mid x_i) = \begin{cases} g_1(x_i), & \text{if } x_i \geq x_0, \\ g_0(x_i), & \text{if } x_i < x_0, \end{cases} \qquad \text{where } g_1(x_0) \neq g_0(x_0). \tag{10.3}$$

Let us assume that $g_1(x_0) > g_0(x_0)$ without loss of generality. Thus, surpassing the cutoff makes treatment more likely. Figure 10.3 illustrates: the treatment probability is strictly between zero and one on both sides, and jumps up at the cutoff. The sharp design is the special case $g_0 = 0$, $g_1 = 1$.

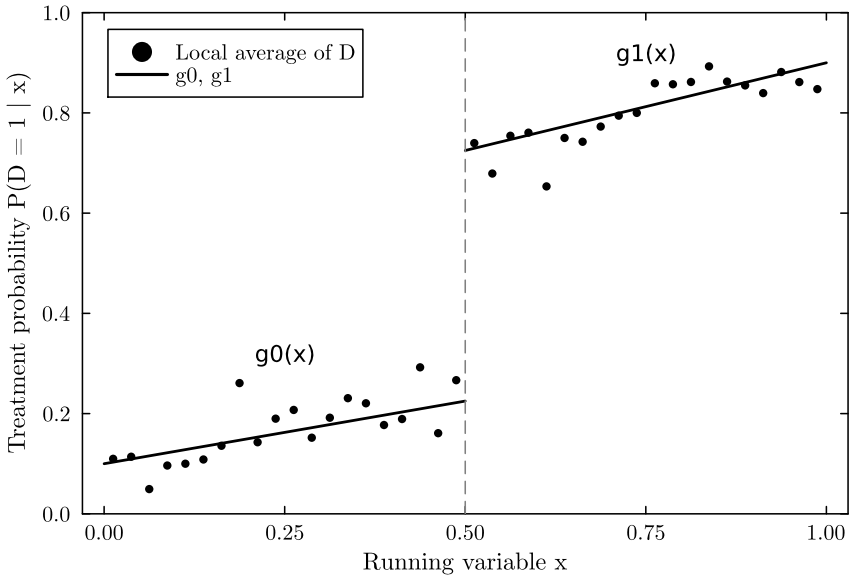


Figure 10.3: The first stage of a fuzzy RD. Dots are local averages of the treatment indicator in bins of the running variable. The treatment probability jumps from $g_0(x_0)$ to $g_1(x_0)$ at the cutoff, but treatment is not deterministic on either side.

10.3.2 Fuzzy RD as 2SLS

Denote $T_i = \mathbf{1}(x_i \geq x_0)$ as the indicator of whether observation i passes the cutoff. Then we can naturally write fuzzy RD as a two stage least squares problem. The treatment D_i is the endogenous variable, and the cutoff indicator T_i is the instrument:

- **First stage:** regress the treatment D_i on the cutoff indicator T_i (and the smoothing function);
- **Second stage:** regress the outcome Y_i on the first stage fitted value $\hat{D}_i$ (and the smoothing function).

The smoothing function f in the running variable should be included in *both* stages: the exclusion restriction for T_i is only credible after controlling smoothly for x_i. Intuitively, the reduced form jump in Y at the cutoff is scaled up by the first stage jump in D: the fuzzy RD estimand is the ratio of the outcome discontinuity to the treatment discontinuity, a Wald estimator localized at x_0. We will make this precise in Section 10.4.

Implementation in Stata

It is very simple to implement RD in Stata. Packages such as `rdrobust` automate the whole pipeline: they implement bias corrected confidence intervals with the optimal bandwidth of Calonico et al. (2014), and you can also try the optimal bandwidth of Imbens and Kalyanaraman (2012). The package chooses the bandwidth h by minimizing an asymptotic mean squared error criterion, fits local polynomials on each side, and corrects the confidence interval for the smoothing bias. In this course's practice example we will also build the estimator by hand with `regress` and `ivregress` to see the mechanics.

10.4 Non-Parametric Identification of RD

We have already introduced how to implement the RD method, and intuitively discussed its identification source. But what kind of causal effect are we identifying? And what exactly are its identification assumptions? Let us go to a classic study in RDD: Hahn et al. (2001), henceforth HTV. Do not say that you understand RDD if you never read this paper.

10.4.1 Setup and Notation

Notation change in this section

To stay close to HTV, this section switches notation: the *treatment* is now denoted x_i (binary), and the *running variable* is denoted z_i with cutoff z_0. Do not confuse this with Sections 10.2 and 10.3, where x_i was the running variable.

Denote y_{1i}, y_{0i} as the potential outcomes and $x_i \in \{0, 1\}$ as the treatment. Write the observed outcome as

$$y_i = \alpha_i + x_i \beta_i, \tag{10.4}$$

so that $\alpha_i \equiv y_{0i}$ is the untreated potential outcome and $\beta_i \equiv y_{1i} - y_{0i}$ is the individual treatment effect. Both are allowed to vary freely across individuals.

Assume that we have a running variable z_i. In a *sharp* design, $x_i = f(z_i)$ is a deterministic function of z_i, discontinuous at z_0. In a *fuzzy* design, $P(x_i = 1 \mid z_i) = f(z_i)$ is discontinuous at z_0. The basic requirement, covering both cases, is that treatment

take up actually jumps at the cutoff.

Assumption 10.4.1 (RD, in HTV)**.** (i) The limits $x^+ \equiv \lim_{z \to z_0^+} E[x_i \mid z_i = z]$ and $x^- \equiv \lim_{z \to z_0^-} E[x_i \mid z_i = z]$ exist. (ii) $x^+ \neq x^-$.

In the sharp design $x^+ = 1$ and $x^- = 0$; in the fuzzy design $x^+ = g_1(z_0)$ and $x^- = g_0(z_0)$. Assumption 10.4.1 plays the role of the first stage relevance condition in IV. Define the corresponding outcome limits

$$y^+ \equiv \lim_{z \to z_0^+} E[y_i \mid z_i = z], \qquad y^- \equiv \lim_{z \to z_0^-} E[y_i \mid z_i = z].$$

10.4.2 Constant Treatment Effects

First, consider the simple case of constant treatment effects: $\beta_i = \beta$ across individuals. What do we need beyond Assumption 10.4.1? We need the mean *untreated* potential outcome to be continuous at the cutoff. That is, the mean of all other confounders must be continuous at the cutoff: nothing except the treatment probability jumps at z_0.

Assumption 10.4.2 (A1 in HTV)**.** $E[\alpha_i \mid z_i = z]$ is continuous in z at z_0.

We can then prove that β is non-parametrically identified.

Theorem 10.4.1 (Theorem 1 in HTV)**.** *Suppose that β_i is fixed at β. Further suppose that Assumptions 10.4.1 and 10.4.2 hold. We then have*

$$\beta = \frac{y^+ - y^-}{x^+ - x^-}.$$

Using an IV style method, we can pin down the treatment effect: the estimand is the jump in the outcome divided by the jump in the treatment probability, exactly the localized Wald ratio anticipated in Section 10.3. In the sharp case the denominator is 1 and the estimand reduces to the outcome jump $y^+ - y^-$ itself. You will derive this theorem step by step in Homework Problem 1.

10.4.3 Heterogeneous Treatment Effects

Next, we go to the more complicated heterogeneous treatment effect case, where β_i varies across individuals. We need one more assumption: not only must α be continuous

at z_0, but also β.

Assumption 10.4.3 (A2 in HTV)**.** $E[\beta_i \mid z_i = z]$ is continuous in z at z_0.

It is worth pausing on how Assumptions 10.4.2 and 10.4.3 differ, because they are conceptually distinct. A1 is an *exogeneity* assumption: there is no systematic difference in the untreated outcome y_{0i} around the cutoff. Violation example: students with very high ability can precisely control their scores to land just above the cutoff line. Then the just above group has better y_{0i} than the just below group, and the comparison is contaminated.

A2 is a *no treatment effect sorting* assumption: there is no systematic difference in the gain $y_{1i} - y_{0i}$ around the cutoff. Violation example: students with a high return to treatment are more likely to select into treatment near the cutoff. Then the treated units just above the cutoff have unusually large β_i. Then we have the following theorem.

Theorem 10.4.2 (Theorem 2 in HTV)**.** *Suppose that x_i is independent of β_i conditional on z_i near z_0. Further suppose that Assumptions 10.4.1, 10.4.2, and 10.4.3 hold. We then have*

$$E[\beta_i \mid z_i = z_0] = \frac{y^+ - y^-}{x^+ - x^-}.$$

Theorem 10.4.2 tells us that under heterogeneous treatment effects, if other confounding factors are continuous at the cutoff (A1), and there is no sorting on returns at the cutoff (A2 together with the local independence condition), then we can identify the average treatment effect for individuals around the cutoff. Since treatment is locally independent of the gains, this average over everyone at the cutoff coincides with the average for the treated there, so we can identify the *ATT for individuals around the cutoff*.

However, "no sorting" is a strong assumption under fuzzy RD. When take up is voluntary, individuals of course choose treatment based on how much they can benefit from it. That is exactly what the Roy model of Chapter 6 tells us: selection operates on gains. In a fuzzy design where crossing the cutoff merely makes treatment cheaper or more likely, the people who take treatment just above the cutoff are precisely those with high β_i, and the local independence condition fails.

10.4.4 Dropping No Sorting: a LATE Result

Let us see what happens if we drop the no sorting condition. We invoke a set of assumptions similar to Imbens and Angrist (1994) on LATE. Write $x_i(z)$ for the treatment that individual i would take if the running variable were externally set to z.

Assumption 10.4.4 (A3 in HTV)**.** (i) $(\beta_i, x_i(z))$ is jointly independent of z_i near z_0. (ii) There exists $\epsilon > 0$ such that $x_i(z_0 + e) \geq x_i(z_0 - e)$ for all $0 < e < \epsilon$.

Part (i) is the local exclusion restriction: given the choice mapping $x_i(\cdot)$, the treatment effect β_i is independent of the realized position of z_i near z_0. The running variable z can only affect y through changing the treatment x. In the admission example: test scores only affect wages through changing whether you can be admitted to PKU. Part (ii) is monotonicity in a small neighborhood around the cutoff: crossing the threshold never pushes anyone *out* of treatment. These are exactly the LATE conditions of Chapter 5, localized at z_0.

Theorem 10.4.3 (Theorem 3 in HTV)**.** *Suppose that Assumptions 10.4.1, 10.4.2, and 10.4.4 hold. We then have*

$$\lim_{e \to 0^+} E\big[\beta_i \,\big|\, x_i(z_0 + e) - x_i(z_0 - e) = 1\big] = \frac{y^+ - y^-}{x^+ - x^-}.$$

Theorem 10.4.3 says that we can identify a LATE under a set of assumptions similar to Imbens and Angrist (1994). Note that this LATE is "local" in two distinct senses. First, it is local in *behavior*: it averages β_i only over individuals who change their treatment choice when the running variable crosses the cutoff, the compliers. Second, it is local in *position*: it averages only over individuals with z_i around the cutoff z_0. The estimand is the average effect for compliers around the cutoff, a doubly local parameter. Whether it is the parameter you care about depends on the question, but it is what the design honestly delivers.

What RD Identifies

All three HTV theorems share the same sample statistic, the localized Wald ratio $(y^+ - y^-)/(x^+ - x^-)$. What changes is its interpretation:

- constant effects (Theorem 10.4.1): the common effect β;
- heterogeneous effects with no sorting (Theorem 10.4.2): the ATT for individuals at the cutoff;

- heterogeneous effects with self selection (Theorem 10.4.3): the LATE for compliers at the cutoff.

Sharp RD is matching at the cutoff; fuzzy RD is IV at the cutoff.

10.4.5 Validating the Design in Practice

From this analysis of the identification of RD, we can derive what conditions we have to validate in an empirical application.

First, we need to check the *existence of the discontinuity* (Assumption 10.4.1). Draw the figure with the running variable on the x axis and the treatment on the y axis; draw the figure with the running variable on the x axis and the outcome on the y axis. Visually detect the discontinuity, as in Figures 10.2 and 10.3. If the first stage jump is not visible to the naked eye, no amount of econometrics will rescue the design.

Second, implement a *balance test* for samples just below and just above the cutoff (Assumptions 10.4.2 and 10.4.3). Other variables and confounders, especially predetermined characteristics, should be similar or continuous around the cutoff. The logic is that of a placebo: a variable determined before treatment cannot jump at the cutoff unless there is a selection there.

Additionally, check the *density* of the sample around the cutoff. Make sure there is no bunching to either one side of it. Good students should not be able to control their scores to land just a little above the threshold; if they can, the just above group is selected and A1 fails. A histogram of the running variable with a spike just above the cutoff and a trough just below is the classic red flag. McCrary (2008) proposes a formal statistical test for this.

The RD Validation Checklist

1. Plot treatment against the running variable: the first stage jump must be visible.
2. Plot the outcome against the running variable: inspect the reduced form jump.
3. Balance test: plot and test predetermined covariates around the cutoff; they must not jump.
4. Density test: check the histogram of the running variable for bunching at the

cutoff. A formal version of this test is developed by McCrary (2008).

10.5 Extension of RDD: Regression Kink Design

10.5.1 From Jumps to Kinks

An interesting extension of RDD is the *Regression Kink Design* (RKD). Rather than using a discontinuity in the *level* of treatment, we employ a *kink*: the jump is no longer in the level, but in the slope. Or we say, the treatment (or treatment probability) has a discontinuity in its first derivative with respect to the running variable, a second order discontinuity.

10.5.2 The Leading Example: Unemployment Insurance

Consider the example from Card et al. (2015) and Card et al. (2017). In many countries, workers receive compensation when they are unemployed, called unemployment insurance benefit (UI). The amount of UI depends on the wage of your last job. If your last wage is too low, there is a minimum benefit level. There is also a maximum value for UI: Bill Gates will not get billions once he is unemployed.

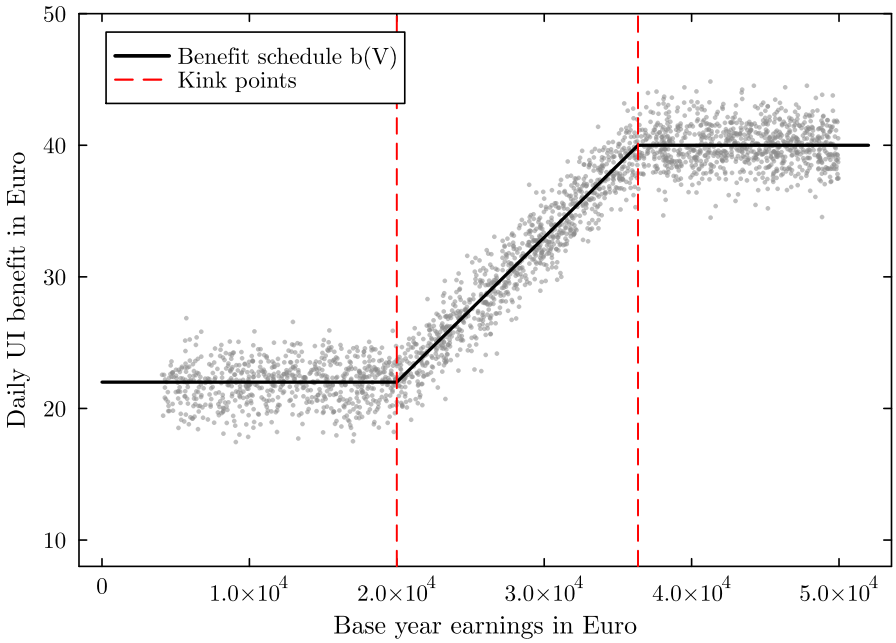


Figure 10.4: The UI benefit schedule in the style of Card et al. (2017), Austria 2004, replicated on simulated data. The daily UI benefit is flat at the minimum for low base year earnings, rises proportionally in the middle range, and is capped at the maximum. Two kinks are noticeable: at the minimum and at the maximum.

Figure 10.4 shows the simulated benefit schedule for Austria in 2004 based on Card et al. (2017): benefits are flat at the minimum, then rise linearly with base year earnings, then are capped at the maximum. Two kinks are noticeable, one at the minimum and one at the maximum. At each kink the *level* of benefits is continuous, but the *slope* changes abruptly.

A controversial issue is that overly generous UI can incentivize workers not to search for new jobs, the classic moral hazard concern. It is therefore important to investigate the relation between the UI benefit B and unemployment duration Y. But B is not randomly assigned: it is a function of the last wage, which also affects job finding directly. The kinks provide the identifying variation.

10.5.3 Setup

We aim to investigate the impact of unemployment benefit B on unemployment duration Y. Denote V as the wage of the last job, the running variable, and U as an error

term capturing all unobserved heterogeneity. We have the outcome function

$$Y = y(B, V, U), \tag{10.5}$$

so duration may depend on the benefit B, directly on the last wage V, and on unobservables U. In a *sharp kink design*, B is a deterministic function of V:

$$B = b(V), \qquad \text{with a slope change at } V = 0,$$

where we normalize the kink location to $V = 0$ to simplify the notation. Note the contrast with sharp RD: there the treatment level jumps at the cutoff; here the treatment level is continuous but its derivative jumps.

10.5.4 Identification Assumptions

Four assumptions deliver identification (Card et al., 2015).

Assumption 10.5.1 (Regularity)**.** (i) U is bounded. (ii) y is continuous and partially differentiable with respect to b and v, and the partial derivative $y_b(b, v, u)$ is continuous.

Assumption 10.5.2 (Exclusion)**.** $y_v(b, v, u)$ is continuous around the kink $v = 0$: the kink exists only for the schedule $b(v)$, not for the direct effect of v on y.

Assumption 10.5.3 (Kink existence)**.** The treatment assignment rule $b(v)$ is known, continuous, and has a kink at $v = 0$.

Assumption 10.5.4 (No kink for confounders)**.** The conditional density $f_{V|U}(v)$ and its partial derivative with respect to v are continuous around the kink $v = 0$.

Assumption 10.5.2 is the RKD analogue of the exclusion restriction: the last wage may affect duration directly, but that direct effect must be *smooth* through the kink point, so any slope change in outcomes at the kink is attributable to the benefit schedule. Assumption 10.5.4 is the analogue of the no sorting condition: the distribution of unobserved types must evolve smoothly through the kink, ruling out workers manipulating their base wage to sit exactly at the kink.

10.5.5 The Identification Result

Proposition 10.5.1 (Proposition 1 in Card et al., 2015)**.** *In a valid sharp RKD, that is, when Assumptions 10.5.1 to 10.5.4 hold: (a) $P(U \leq u \mid V = v)$ is continuously differentiable in v at $v = 0$ for all $u \in I_U$, where I_U is the bounded support of U; (b)*

$$E\big[y_b(b_0, 0, U) \,\big|\, V = 0\big] = \frac{\lim\limits_{v_0 \to 0^+} \left.\dfrac{dE[Y \mid V = v]}{dv}\right|_{v=v_0} - \lim\limits_{v_0 \to 0^-} \left.\dfrac{dE[Y \mid V = v]}{dv}\right|_{v=v_0}}{\lim\limits_{v_0 \to 0^+} \left.\dfrac{db(v)}{dv}\right|_{v=v_0} - \lim\limits_{v_0 \to 0^-} \left.\dfrac{db(v)}{dv}\right|_{v=v_0}},$$

where $b_0 = b(0)$ is the benefit level at the kink.

Sharp RKD is dividing the slope change of $E[Y \mid V]$ by the slope change of $b(V)$. On the contrary, RDD divides level by level. Sharp RKD identifies the average marginal effect of the treatment, $E[y_b(b_0, 0, U) \mid V = 0]$, for individuals at the kink: that is, the ATT for individuals with $B = b_0$ and $V = 0$.

The intuition is as follows. A change in the slope of the treatment schedule results in a change in the slope of the average outcome. If there is no change of slope for unobserved confounders (Assumptions 10.5.2 and 10.5.4), then we can attribute all the slope change in outcomes to the slope change in treatment. The ratio of the two slope changes is the marginal effect of the treatment, just as the ratio of two level jumps is the treatment effect in RDD.

10.5.6 Fuzzy RKD

What about the fuzzy case, where the schedule is not deterministic (actual benefits deviate from the formula because of measurement error, take up decisions, or unobserved eligibility rules)? The result is much more complicated, but with no surprising intuition. In a fuzzy RKD, we identify a LATE-style parameter: a weighted average of marginal effects for individuals who have UI slope changes at the kink. The larger your own slope change at the kink, the larger weight you receive in the average. Of course, we have to invoke a monotonicity assumption, exactly parallel to the fuzzy RD case: the kink must shift everyone's benefit slope in the same direction.

RDD versus RKD

- **RDD:** treatment level jumps at the cutoff. Estimand = outcome jump divided by treatment jump. Sharp: ATT at the cutoff. Fuzzy: LATE for compliers at the cutoff.
- **RKD:** treatment slope changes at the kink. Estimand = outcome slope change divided by treatment slope change. Sharp: average marginal effect at the kink. Fuzzy: weighted LATE for individuals whose treatment slope changes at the kink.

10.6 Application of RD: He, Wang, and Zhang (2020)

At the end of this chapter, we introduce an application of the RD method, He et al. (2020), published in the *Quarterly Journal of Economics*. The paper estimates the effect of environmental regulation on firm productivity in China, and the basic idea is very interesting.

Research question and design. How much does environmental regulation cost in terms of firm productivity? To improve surface water quality, the Chinese central government installed several hundred state controlled water quality monitoring stations along major rivers, and used the water quality readings to evaluate local government officials for promotion. The key institutional detail is physical: a monitoring station only captures emissions from its *upstream* regions. Water flows one way, so a polluting firm just upstream of a station affects the reading, while an identical firm just downstream does not. Local government officials, whose careers depend on the readings, are therefore incentivized to enforce tighter environmental standards on firms immediately upstream of a monitoring station rather than those immediately downstream. This gives a natural RDD setting: the running variable is the firm's position along the river relative to the station, the cutoff is the station itself, and crossing from downstream to upstream discontinuously raises regulatory stringency. It is a spatial regression discontinuity, with the same logic as the exam score design but with geography as the running variable.

Main results. Using novel firm level geocoded emission and production data, the authors find that immediate upstream polluters have more than 24 percent lower total

factor productivity (TFP) and more than 57 percent lower chemical oxygen demand (COD) emissions than their immediate downstream counterparts. The TFP discontinuity does not exist in nonpolluting industries, which face no differential regulation; it only emerged after 2003, when the government explicitly linked water quality readings to political promotion; and it is predominantly driven by prefectural cities with career driven leaders. These placebo and heterogeneity patterns tie the discontinuity to the regulation channel rather than to endogenous location choices of stations or firms.

Interpretation. Combining the TFP and emission estimates, a back of the envelope calculation indicates that China's water regulation efforts between 2000 and 2007 were associated with an economic cost of more than 800 billion Chinese yuan. Beyond the cost accounting, the paper documents a general political economy lesson: when the center rewards local officials on a measurable proxy (readings at monitoring stations) rather than the true objective (water quality), enforcement gets distorted toward what is measured, generating large spatial inequities in regulatory burden. For our purposes, the paper is a model of how to find a credible discontinuity inside an institution and how to defend it with placebo tests.

10.7 Conclusion

When you have a discontinuity in treatment, you can use RDD. The chapter's conceptual map is compact:

- **Sharp RDD is matching.** It uses samples around the cutoff, and it identifies the ATT for individuals around the cutoff.
- **Fuzzy RDD is IV.** It uses the cutoff indicator as an instrument, and it identifies the LATE for compliers around the cutoff.
- When you have a discontinuity in the treatment *slope*, you can use RKD, which identifies the corresponding ATT and LATE-style parameters in sharp and fuzzy settings, respectively.

Practical Tips for RD in Applied Work

In practice, remember the following tips:

1. Do not use high order polynomials as smoothing functions (Gelman and

Imbens, 2019).

2. A common way is to use local linear regression around the cutoff.
3. Use packages in Stata (such as `rdrobust`) to obtain the optimal bandwidth and bias corrected inference (Calonico et al., 2014; Imbens and Kalyanaraman, 2012).
4. Implement balance tests both visually and statistically to validate your assumptions, and check the density of the running variable for bunching.

Homework Problems

1. **Deriving Theorem 1 in HTV (2001).**

   Consider the setup of Section 10.4: $y_i = \alpha_i + x_i\beta_i$ with a constant treatment effect $\beta_i = \beta$, a running variable z_i with cutoff z_0, and Assumptions 10.4.1 and 10.4.2 in force.

   (a) Write $E[y_i \mid z_i = z]$ in terms of $E[\alpha_i \mid z_i = z]$ and $E[x_i \mid z_i = z]$, and take limits from above and below z_0 to express y^+ and y^-.

   (b) Use Assumption 10.4.2 to show that the α terms cancel, and conclude that

   $$\beta = \frac{y^+ - y^-}{x^+ - x^-}.$$

   (c) Show that in the sharp design the formula reduces to $\beta = y^+ - y^-$, and explain in one sentence why no first stage estimation is needed there.

   (d) Where exactly did you use part (ii) of Assumption 10.4.1? What goes wrong with the formula if it fails?

2. **A1 versus A2.**

   Answer in words; do not use math.

   (a) Explain the difference between Assumption A1 and Assumption A2 in HTV (2001): what kind of systematic difference around the cutoff does each one rule out?

   (b) Give one concrete violation example for each assumption, in the setting of a university admission cutoff.

(c) Why is the no sorting condition behind Theorem 10.4.2 considered particularly strong in a *fuzzy* design? Connect your answer to the Roy model.

(d) Which of the two assumptions can be indirectly checked with a density test of the running variable, and why?

3. **A Sharp RKD Computation.**

An unemployment insurance system pays a daily benefit equal to 80 percent of the previous daily wage, up to a cap: every worker whose previous wage exceeds the cap threshold receives the same maximum benefit. Normalize the running variable V to be the previous wage minus the cap threshold, so the kink is at $V = 0$. The benefit schedule is

$$b(V) = \begin{cases} b_0 + 0.8\,V, & V < 0, \\ b_0, & V \geq 0. \end{cases}$$

Using administrative data, you estimate the slope of expected unemployment duration in the previous wage on each side of the kink:

$$\lim_{v \to 0^-} \frac{dE[Y \mid V = v]}{dv} = 1.5, \qquad \lim_{v \to 0^+} \frac{dE[Y \mid V = v]}{dv} = 0.3,$$

where Y is measured in days and wages in RMB.

(a) Compute the slope change of the benefit schedule and the slope change of expected duration at the kink.

(b) Apply Proposition 10.5.1 to compute the sharp RKD estimand, and state its units.

(c) Interpret the sign and magnitude of your answer in terms of the moral hazard debate about UI generosity.

(d) Suppose you discover that the density of V has a visible kink at $V = 0$ (for example, employers bunch wages exactly at the cap threshold). Which assumption fails, and why does your number in (b) lose its causal interpretation?

Practice Example: Sharp and Fuzzy RD in a Simulation

In this practice example we simulate a sharp and a fuzzy RD dataset, estimate the treatment effect with local linear regression built by hand, implement the fuzzy design as 2SLS, and run the full validation checklist of Section 10.4.5.

Data generating process, sharp design. We simulate $n = 5000$ individuals with a running variable $x_i \sim \text{Uniform}(-1, 1)$ and cutoff $x_0 = 0$. Treatment is deterministic, $D_i = \mathbf{1}(x_i \geq 0)$. The outcome is

$$Y_i = 1 + 1.5\, x_i + 0.8\, x_i^2 + 2\, D_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 0.5^2),$$

so the true treatment effect at the cutoff is $\rho = 2$ and the conditional mean has mild curvature. A predetermined covariate $w_i = 0.5\, x_i + \nu_i$ with $\nu_i \sim \mathcal{N}(0, 1)$ is also generated; it depends on the running variable smoothly but does not jump at the cutoff, so it should pass the balance test.

Data generating process, fuzzy design. Same running variable, but treatment is stochastic:

$$P(D_i = 1 \mid x_i) = \begin{cases} 0.2 + 0.2\, x_i, & x_i < 0, \\ 0.7 + 0.2\, x_i, & x_i \geq 0, \end{cases}$$

so the first stage jump at the cutoff is 0.5. The outcome is $Y_i = 1 + 1.5\, x_i + 2\, D_i + \epsilon_i$ with the same error distribution, and the true effect is again 2. Take up is independent of ϵ_i given x_i and the effect is constant, so Assumptions 10.4.1 and 10.4.2 of Theorem 10.4.1 hold by construction.

Analysis steps. The do-file (`src/chapter 10/ch10_rd.do`) performs:

1. Simulate the sharp dataset and save it; produce a binned scatter of Y against x to visualize the jump.
2. Estimate the sharp design by global polynomials of increasing order, and observe how the estimate moves with the polynomial degree, illustrating the sensitivity criticized by Gelman and Imbens (2019).
3. Estimate the sharp design by local linear regression, interacting the centered running variable with the treatment side, for bandwidths $h = 0.50, 0.25, 0.10$;

compare the estimates with the truth $\rho = 2$.

4. Run the balance test: local linear regression of the predetermined covariate w at the cutoff; verify there is no jump.
5. Run the density check: compare the number of observations in narrow bins on each side of the cutoff.
6. Simulate the fuzzy dataset and save it; verify the first stage jump of about 0.5 in the treatment probability.
7. Estimate the fuzzy design by 2SLS with `ivregress`, instrumenting D_i with the cutoff indicator $T_i = \mathbf{1}(x_i \geq 0)$ and controlling for the running variable on both sides, within a bandwidth; compare with the naive OLS of Y on D, which is contaminated by selection on the running variable.
8. Key lessons summary.

Key lessons.

- The RD estimate is a local object: only observations near the cutoff carry information, and the bandwidth governs the bias variance tradeoff directly.
- Global high order polynomial fits are unstable in the polynomial degree, while local linear estimates are stable across reasonable bandwidths.
- The fuzzy design is exactly IV: the reduced form jump divided by the first stage jump, and `ivregress` with the cutoff indicator as the instrument reproduces this ratio.
- Balance and density checks are cheap and should be reported in any RD application.

The do-file is located at `src/chapter 10/ch10_rd.do` and the simulated datasets at `data/chapter 10/`.

Part IV

Other Topics

Chapter 11

Standard Error Issues

11.1 Introduction

Inference is important in practice. Estimation takes us from data to a target of interest, but a point estimate alone answers only half of the question. How accurate is our estimate? How confident are we in our results? The other half of the answer is the standard error, and this chapter is about the situations where the standard error you type out of a regression command is the wrong one.

In the traditional inference of econometrics, two assumptions are doing the work behind the familiar formulas:

1. Uncertainty comes from *random sampling*, and the asymptotics let the sample size go to infinity, $n \to \infty$;
2. The sample is *i.i.d.*: observations are drawn independently, with no correlations across them.

What if these two assumptions are violated? In this lecture we consider two such cases.

First, sometimes n is naturally limited. If your observations are the provinces of China, there are only about thirty of them, and no asymptotic thought experiment can manufacture more. When the sample is a large fraction of the population, another type of uncertainty becomes important: *design-based uncertainty*, the uncertainty coming from the random assignment of treatment rather than from the drawing of the sample. Section 11.2 develops this idea following Abadie et al. (2020).

Second, the i.i.d. assumption may fail because errors are *clustered*: students in the same class, workers in the same firm, residents of the same city share common shocks, and their errors are correlated. We have to incorporate this structure in inference, or the reported standard errors can be badly wrong. Sections 11.3 and 11.4 develop the clustering problem: the Moulton factor that quantifies the bias, the estimators that fix it, and, following Abadie et al. (2023), the harder question of at which level one should cluster.

Angrist and Pischke (2009) call this family of problems "nonstandard standard error issues," and the name is apt: the point estimates are fine and everything that goes wrong goes wrong in the standard error.

11.2 Design-based Uncertainty

11.2.1 What Is Inference?

In the usual case, when we talk about inference, what is that? It is useful to fix the vocabulary first. We have a target parameter, the *estimand* β: this is the **target**. We want to recover it using an *estimator* $\hat{\beta}$, a rule applied to a sample from the population: this is the **method**. Applying the rule to our particular sample gives a number, the *estimate*, say $\hat{\beta} = 0.5$: this is the **result**. The whole process is called estimation, or statistical inference: the **process**.

Usually, we consider *sampling-based* uncertainty. Each time you draw a new sample from the population, it gives you a new estimate from your estimation process. When the sample changes, your estimation result changes. Uncertainty comes from the sampling process, and this is the uncertainty your standard error is supposed to measure. But is this the only uncertainty in empirical research? Today we are going to introduce a second source.

11.2.2 Two Sources of Uncertainty

Design-based uncertainty, introduced by Abadie et al. (2020) (henceforth AAIW), is the uncertainty coming from the *treatment assignment*. The treatment X_i is no longer considered fixed: in some states of the world person 1 is treated, in others person 1 is not treated. The potential outcome you observe is different when the treatment is randomly

changed. As we will show, this perspective helps you understand the uncertainty of estimation when the sample is a non-negligible fraction of the population.

To visually explain the difference between traditional sampling-based uncertainty and design-based uncertainty, let us take a look at two tables from AAIW. Let $R_i \in \{0, 1\}$ be an indicator of whether observation i is included in the sample.

Table 11.1: Sampling-based uncertainty (✓ is observed, ? is missing), adapted from Table I in Abadie et al. (2020). Y_i is the outcome, Z_i an attribute, R_i the sampling indicator.

	Actual Sample			Alt. Sample I			Alt. Sample II			
Unit	Y_i	Z_i	R_i	Y_i	Z_i	R_i	Y_i	Z_i	R_i	$\cdots$
1	✓	✓	1	?	?	0	?	?	0	$\cdots$
2	?	?	0	?	?	0	?	?	0	$\cdots$
3	?	?	0	✓	✓	1	✓	✓	1	$\cdots$
4	?	?	0	✓	✓	1	?	?	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$
n	✓	✓	1	?	?	0	?	?	0	$\cdots$

Table 11.2: Design-based uncertainty (✓ is observed, ? is missing), adapted from Table II in Abadie et al. (2020). $Y_i^*(1), Y_i^*(0)$ are potential outcomes, X_i the treatment.

	Actual Sample			Alt. Sample I			Alt. Sample II			
Unit	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$\cdots$
1	✓	?	1	✓	?	1	?	✓	0	$\cdots$
2	?	✓	0	?	✓	0	?	✓	0	$\cdots$
3	?	✓	0	✓	?	1	✓	?	1	$\cdots$
4	?	✓	0	✓	?	1	?	✓	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$
n	✓	?	1	?	✓	0	?	✓	0	$\cdots$

The two tables encode two different thought experiments:

- **Sampling-based uncertainty** (Table 11.1): treatment is fixed, and the sampled observations are random. For non-sampled individuals, we cannot observe

anything, outcome and attributes alike. The source of uncertainty is that in each sample we have *different observations*.

- **Design-based uncertainty** (Table 11.2): treatment is random, and the set of observations is fixed, for example all provinces of China. For each individual, we only observe the potential outcome in the realized treatment status, never the counterfactual one. The source of uncertainty is that in each sample we have a *different treatment status for each individual*.

Next, AAIW construct a simple model and make the following four points, which we develop one by one:

1. show how design-based uncertainty affects the variance of the regression estimator;
2. show that the White estimator remains conservative when we consider design-based uncertainty;
3. derive a finite population correction for the White estimator;
4. discuss the two sources of uncertainty and their link to external and internal validity.

11.2.3 A Finite Population Framework

Assume that we have a **finite** population of size n. We randomly sample N observations from the n, with $R_i \in \{0, 1\}$ the indicator of whether i is sampled. There is a random binary treatment regressor X_i: in the population, n_1 units are treated and n_0 are not; in the sample, N_1 are treated and N_0 are not. Each unit has potential outcomes $Y_i^*(1)$ and $Y_i^*(0)$, and we observe

$$Y_i = Y_i^*(X_i) = \begin{cases} Y_i^*(1), & \text{if } X_i = 1, \\ Y_i^*(0), & \text{if } X_i = 0. \end{cases}$$

The potential outcomes themselves are assumed to be non-stochastic: all randomness comes from who is sampled ($\mathbf{R}$) and who is treated ($\mathbf{X}$). We use bold letters to represent vectors over the whole population, $(\mathbf{Y}, \mathbf{Y}^*(1), \mathbf{Y}^*(0), \mathbf{R})$.

We define three estimands as our proposed targets.

Definition 11.2.1 (Three estimands, AAIW)**.**

$$\begin{aligned}
\theta^{descr} &= \frac{1}{n_1}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-X_i)Y_i, \\
\theta^{causal,sample} &= \frac{1}{N}\sum_{i=1}^{n} R_i \left(Y_i^*(1) - Y_i^*(0)\right), \\
\theta^{causal} &= \frac{1}{n}\sum_{i=1}^{n} \left(Y_i^*(1) - Y_i^*(0)\right).
\end{aligned}$$

The *descriptive* estimand θ^{descr} is the population mean difference between treated and untreated units. It is free of $\mathbf{R}$ and free of potential outcomes: it compares realized outcomes as they stand, with no causal ambition. The *causal sample* estimand $\theta^{causal,sample}$ is the average causal effect for the current sample. The *causal* estimand θ^{causal} is the average causal effect for the whole population.

Which uncertainty matters for which estimand? The definitions answer directly:

- When estimating θ^{descr}, we do not care about design-based uncertainty: nothing in the target involves the treatment assignment mechanism or potential outcomes.
- When estimating $\theta^{causal,sample}$, we do not care about sampling-based uncertainty: given the current sample, nothing in the target involves the sampling process.
- When estimating θ^{causal}, we care about *both* types of uncertainty.

11.2.4 One Estimator, Three Targets

To estimate these estimands, we use a simple OLS regression of Y_i on X_i in the sample, which for a binary regressor is the difference in means:

$$\hat{\theta} = \frac{1}{N_1}\sum_{i=1}^{n} R_i X_i Y_i - \frac{1}{N_0}\sum_{i=1}^{n} R_i(1-X_i)Y_i. \tag{11.1}$$

Sampling-based uncertainty comes from the randomness of $\mathbf{R}$; design-based uncertainty comes from the randomness of $\mathbf{X}$. We further assume that both sampling and treatment assignment are random.

Proposition 11.2.1 (Unbiasedness for each target, AAIW)**.**

$$\begin{aligned} E[\hat{\theta} \mid \mathbf{X}, N_1, N_0] &= \theta^{descr}, \\ E[\hat{\theta} \mid \mathbf{R}, N_1, N_0] &= \theta^{causal,sample}, \\ E[\hat{\theta} \mid N_1, N_0] &= \theta^{causal}. \end{aligned}$$

Read the conditioning sets carefully: conditioning on (that is, fixing) the treatment vector, $\hat{\theta}$ is unbiased for the descriptive estimand; conditioning on the sampling vector, $\hat{\theta}$ is unbiased for the causal sample estimand; conditioning on neither, $\hat{\theta}$ is unbiased for the population causal estimand. The same number on your screen is an unbiased estimate of three different targets, and which uncertainty you should report depends on which target you claim.

11.2.5 Three Variances

We define the population variances as follows:

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i^*(x) - \frac{1}{n} \sum_{j=1}^{n} Y_j^*(x) \right)^2, \quad \text{for } x = 0, 1,$$

$$S_\theta^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i^*(1) - Y_i^*(0) - \frac{1}{n} \sum_{j=1}^{n} \big(Y_j^*(1) - Y_j^*(0) \big) \right)^2.$$

S_x^2 is the variance of potential outcomes for the population, and S_θ^2 is the variance of the *treatment effect* for the population: it is zero exactly when the effect is homogeneous.

Based on these population variances, we can derive three variances for $\hat{\theta}$.

Proposition 11.2.2 (Three variances, AAIW)**.**

$$\begin{aligned} V^{total}(N_1, N_0, n_1, n_0) &= \text{var}(\hat{\theta} \mid N_1, N_0) \\ &= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{n_0 + n_1}, \qquad (11.2) \\ V^{sampling}(N_1, N_0, n_1, n_0) &= E[\text{var}(\hat{\theta} \mid \mathbf{X}, N_1, N_0) \mid N_1, N_0] \\ &= \frac{S_1^2}{N_1}\left(1 - \frac{N_1}{n_1}\right) + \frac{S_0^2}{N_0}\left(1 - \frac{N_0}{n_0}\right), \qquad (11.3) \\ V^{design}(N_1, N_0, n_1, n_0) &= E[\text{var}(\hat{\theta} \mid \mathbf{R}, N_1, N_0) \mid N_1, N_0] \\ &= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\theta^2}{N_0 + N_1}. \qquad (11.4) \end{aligned}$$

The interpretation mirrors Proposition 11.2.1:

- V^{total} is the total variance, considering both sampling-based and design-based uncertainty. It is *the variance we want to capture in inference for the causal estimator* of θ^{causal}.
- $V^{sampling}$ is the variance from only sampling-based uncertainty, obtained by conditioning on the treatment assignment. It is the variance for inference on the *descriptive* estimand.
- V^{design} is the variance from only design-based uncertainty, obtained by conditioning on the current sample. It is the variance for inference on the *causal sample* estimand.

Traditionally, we only consider sampling uncertainty, which makes $V^{sampling}$ our default variance. Now let us analyze these formulas one by one.

Point 1: the two partial variances cannot be ranked. Generally, $V^{sampling}$ and V^{design} cannot be ranked; their comparison depends on the sampling rate N/n. A very large sampling rate makes the finite population correction factors $(1 - N_1/n_1)$ and $(1 - N_0/n_0)$ small, and hence a very small $V^{sampling}$: if you have sampled nearly the whole population, redrawing the sample changes almost nothing, while the randomness of treatment assignment remains fully alive.

Point 2: with an infinite population, traditional variance is fine. When $n \to \infty$

with N fixed, the correction factors go to one and the last term of (11.2) vanishes, so

$$V^{sampling} = V^{total}.$$

If the population is infinite, then design-based uncertainty is ignorable, and traditional inference for the causal estimand, without ever thinking about design-based uncertainty, is fine. This is the precise sense in which the textbook i.i.d. sampling story survives: it is the $N/n \to 0$ limit of the finite population framework.

Point 3: with a finite population, traditional variance overstates. Consider estimating θ^{descr} or θ^{causal} when the population is finite. If we treat it as infinite, both V^{total} and $V^{sampling}$ are overstated:

$$V^{total}(N_1, N_0, \infty, \infty) - V^{total}(N_1, N_0, n_1, n_0) = \frac{S_\theta^2}{n_0 + n_1} \geq 0,$$
$$V^{sampling}(N_1, N_0, \infty, \infty) - V^{sampling}(N_1, N_0, n_1, n_0) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} \geq 0.$$

The standard errors you report are too large: honest, but wastefully conservative.

Point 4: for the sample causal estimand, population size does not matter. Consider estimating $\theta^{causal,sample}$. Formula (11.4) depends only on sample quantities: $V^{design}(N_1, N_0, \infty, \infty) = V^{design}(N_1, N_0, n_1, n_0)$. The relative sample size does not affect the variance conditional on the current sample, so V^{design} is fine even if we think of the population as infinite.

11.2.6 The White Estimator and Its Finite Population Bias

In practice, we usually use the White (heteroskedasticity robust) estimator of the variance matrix, and it is calculated *without* considering design-based uncertainty. In the difference in means case it takes the form

$$\hat{V}^w = \frac{\hat{S}_1^2}{N_1} + \frac{\hat{S}_0^2}{N_0}, \qquad \text{where } \hat{S}_1^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{n} R_i X_i \left(Y_i - \frac{1}{N_1} \sum_{j=1}^{n} R_j X_j Y_j \right)^2,$$

and $\hat{S}_0^2$ is defined analogously.

Proposition 11.2.3 (Bias of the White estimator, AAIW)**.** *$\hat{V}^w$ is unbiased for V^{total}*

when n is infinite. In a finite population, the bias is

$$E[\hat{V}^w \mid N] - V^{total} = \frac{S_\theta^2}{n} \geq 0.$$

The White estimator is therefore *conservative*: it never understates the total variance, and the overstatement, S_θ^2/n, is governed by treatment effect heterogeneity relative to the population size. Note the special case: under homogeneous treatment effects ($S_\theta^2 = 0$) there is no bias at all, even in a tiny population.

We can now see the practical rule. If we ignore design-based uncertainty in inference:

- It is fine if we have a *small sample compared with a massive population*. If you have a CFPS household dataset representing all families in China, the sampling rate is tiny, and the traditional standard error is the right one.
- The positive bias becomes large if we have a *large sample size compared with a limited population*, like a province level regression: you observe essentially all provinces, there is no sampling uncertainty left, yet the traditional formula keeps pricing it in. In this case, traditional variance estimation can be too large and too conservative, because you ignore the fact that you already have a large part of the population.

Fortunately, a bias corrected estimator can be derived, by taking into consideration that you have a large sample relative to a small population and that you have uncertainty in treatment assignment. The derivation of this estimator is technical; read Abadie et al. (2020) if you are interested.

Design-based Uncertainty in a Nutshell

Based on the framework of Abadie et al. (2020), uncertainty has two sources: who is sampled (**R**) and who is treated (**X**). Which one matters depends on the estimand: sampling uncertainty for the descriptive estimand, design uncertainty for the sample causal estimand, both for the population causal estimand. Traditional inference prices only sampling uncertainty with an infinite population; it remains valid when $N/n \approx 0$ and becomes conservative, by S_θ^2/n in the White case, when the sample is a large fraction of a finite population. The number of provinces in China does not grow with your asymptotics. The design-based uncertainty also

helps us to understand how to choose cluster levels when we have clustered standard errors, which we will introduce in the following sections.

11.3 Clustered Standard Errors

11.3.1 Motivating Example: the STAR Experiment

Next, let us consider the clustered standard error issue. Many scholars claim that smaller classes are better for education. What is the impact of class size on students' achievement? This is hard to identify using observational data: better funded schools have both smaller classes and stronger students, the classic selection problem. Project STAR is an RCT designed to answer this question (Krueger, 1999).

The experiment involves about 11,600 children in Tennessee. Kids are randomly assigned to two kinds of classes: (1) small classes with 13 to 17 children; and (2) regular classes with 22 to 25 children. Random assignment removes selection, and we can identify the treatment effect of class size.

One assumption we always make in the regression analysis, however, is i.i.d. sampling. Students in the same class are of course not independently sampled: they share the same classroom, the same teacher, the same peer environment. What will happen if we have correlations at the class, school, or district level? The short answer is: *we may underestimate the standard error*. Let us see why this happens and how to fix it.

11.3.2 Setting and the Intraclass Correlation

Let us go on with the STAR experiment. Consider the following regression for student i in class g:

$$y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}, \tag{11.5}$$

where y_{ig} is the test score, x_g is the class size (randomly assigned), and e_{ig} is the error term. This is a special case in which the regressor x is fixed at the group level g: the same treatment applies to the whole class. Test scores in the same class tend to be correlated: same environment, same teacher, same peers.

Thus, we give up the i.i.d. assumption and assume that for two different students $i \neq j$

in the same class,

$$E[e_{ig}e_{jg}] = \rho_e \sigma_e^2 > 0,$$

where ρ_e is the error intraclass correlation and σ_e^2 is the error variance. A convenient way to generate this structure is to decompose the error into a common class component and an idiosyncratic component:

$$e_{ig} = \nu_g + \eta_{ig}, \qquad \nu_g \perp \eta_{ig},$$

where we assume that ν_g captures all within class correlations (so $\eta_{ig} \perp \eta_{jg}$ for $i \neq j$), and homoskedasticity holds for both ν_g and η_{ig}, with variances σ_ν^2 and σ_η^2. Then we can prove that

$$\rho_e = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}. \tag{11.6}$$

Equation (11.6) is called the *intraclass correlation coefficient*: the intraclass correlation is the share of the common class component in the total error uncertainty. Deriving (11.6) from the setting above is Homework Problem 1.

11.3.3 Bias and the Moulton Factor

How wrong is the conventional standard error when $\rho_e > 0$? Let $V_c(\hat{\beta}_1)$ be the conventional OLS variance, computed as if errors were i.i.d., and let $V(\hat{\beta}_1)$ be the correct variance.

Proposition 11.3.1 (Moulton factor, equal group sizes)**.** *Assume all classes have equal size n and the regressor x_g is fixed within each group. Then*

$$\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)} = 1 + (n-1)\rho_e.$$

We call this ratio the *Moulton factor* (Moulton, 1986). You will derive it in Homework Problem 2. Two comparative statics follow immediately: as n or ρ_e rises, the bias of the conventional variance rises. Larger n means fewer groups for a given total sample, hence less independent information than the observation count suggests. And what happens if $\rho_e = 1$? The conventional variance is only $1/n$ of the correct one (the Moulton factor equals n): students in the same class provide *no* additional information beyond their class mean, so the effective number of observations is the number of classes, not the

number of students.

The previous setting assumes fixed x_g within each group. Let us see the Moulton factor in a more general case, when x_{ig} can vary across i in the same group.

Proposition 11.3.2 (Moulton factor, general case)**.**

$$\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)} = 1 + \left[\frac{V(n_g)}{\bar{n}} + \bar{n} - 1\right] \rho_x \rho_e,$$

where $\bar{n}$ is the average group size, $V(n_g)$ is the variance of group sizes, and ρ_x is the intraclass correlation of x_{ig}.

In general, the bias from within class correlation is larger when:

1. the average group size $\bar{n}$ is larger;
2. the variance of group sizes $V(n_g)$ is larger;
3. the intraclass correlation of the treatment x_{ig}, ρ_x, is larger;
4. the error intraclass correlation ρ_e is larger.

The implications of (3) deserve emphasis. The bias can be very large in the fixed group treatment x_g case, since then $\rho_x = 1$, its maximum. Conversely, there is *no need to cluster anything if the treatment assignment is totally random at the individual level*: then $\rho_x = 0$ and the whole correction term vanishes, no matter how correlated the errors are. The implication of (4) is natural: no bias when $\rho_e = 0$.

11.3.4 Fixing the Bias

Now we know that standard error estimation can be biased when we have correlation within classes. What should we do? Several methods are available:

1. **Moulton correction.** Use the Moulton factor formula of Proposition 11.3.2 to correct the conventional variance directly. Not that good: it relies on the error structure assumptions (homoskedasticity, the additive component model).
2. **Recommended: the Liang and Zeger (1986) clustering estimator.** The cluster robust variance estimator is generally consistent as the number of groups goes to infinity, and requires no model of the within cluster correlation. In Stata, use the option `cluster(`g`)`.

3. **Group level regression.** Run the regression at the group level, $\bar{y}_g = \beta_0 + \beta_1 x_g + \bar{e}_g$, using WLS with group sizes as weights. This has better finite sample properties, but x_g has to be group fixed.

4. **Other methods:** block bootstrap, MLE with a specified error structure, and so on. For more details, please refer to Angrist and Pischke (2009).

Clustered Standard Errors in Stata

The Liang and Zeger estimator is one option away: `reg y x, cluster(classid)`. Remember what the asymptotics need: consistency is in the *number of clusters*, not the number of observations. With 11,600 students in a few hundred classes, the relevant count is the few hundred. With 30 provinces, the cluster robust estimator itself becomes shaky.

11.4 Choosing the Cluster Level

11.4.1 Can We Always Cluster More?

How should we choose the level of clustering? In the STAR experiment, why do we cluster by class and not by boy versus girl, or by racial group? Clustering in more dimensions or at a higher level gives you larger standard errors. Is it then OK to always cluster in more and more dimensions, to be safely conservative? **No.** You can be *too* conservative and overestimate the standard error. Similarly, it is not always good to cluster at a higher and higher level.

The reason connects back to Section 11.2. When you cluster in more and more dimensions, or at a higher and higher level, your *effective* sample size compared with the effective population becomes larger and larger. As Abadie et al. (2020) has shown, a large sampling rate makes traditional (here, clustered) inference overestimate the standard error. For example, suppose you have data on 10,000 firms in 20 provinces. Ten thousand firms can be a very small proportion of all firms in mainland China, so at the firm level the sampling rate is negligible. But when you cluster at the province level, your effective sample becomes 20 provinces out of 31: the effective sampling rate jumps to 20/31, and the infinite population approximation behind the clustered estimator fails badly.

Thus, two issues remain: how to choose the cluster level reasonably, and how to incor-

porate design-based uncertainty. Abadie et al. (2023) consider clustering as a sampling and design problem: the correct cluster level depends on *how you get your sample* and *how you assign your treatment*. It comes from the basic idea of Abadie et al. (2020): you have to consider both sampling-based and design-based uncertainty, and this perspective is the core of the clustering problem.

11.4.2 Three Misconceptions

Abadie et al. (2023) clarify three misconceptions about clustering:

1. *"The need for clustering hinges on the presence of correlation between residuals."* No. The essence is the clustering of *sampling* or of *treatment assignment*. Even if students' scores are correlated within classrooms, there is no need to cluster when sampling and treatment are both totally random: this is exactly the $\rho_x = 0$ lesson of Proposition 11.3.2, now stated at the level of the research design.

2. *"There is no harm in using clustered standard errors when they are not required."* Not true: as we have argued in the previous section, confidence intervals will be unnecessarily conservative, sometimes dramatically so, as the 20/31 example shows.

3. *"Researchers must either fully adjust for clustering by Liang and Zeger, or not adjust at all."* Not really. AAIW propose a new estimator, CCV/TSCB, that corrects for a large effective sampling rate in clustering, interpolating between the two extremes.

11.4.3 The Four Cases

As discussed in Abadie et al. (2020), there are two types of uncertainty we need to consider if our estimation target is the population causal effect, sampling uncertainty and treatment uncertainty. Therefore, we need to take both of them into consideration when we analyze how to cluster the standard error, in a similar way. Here are the empirical suggestions from Abadie et al. (2023), organized by how the sample was drawn and how the treatment was assigned:

1. **Sampling random, treatment random.** Do not cluster! In this case, if the sample additionally represents a large fraction of the population, even the White estimator is too conservative (Abadie et al., 2020).

2. **Sampling random, treatment clustered.** Cluster at the treatment assignment level. In the fuzzy design case (treatment probabilities vary across clusters rather than treatment being fixed within cluster), use the CCV/TSCB estimator. They can be implemented by Stata packages named `ccv` and `tscb`.
3. **Sampling clustered, treatment random.** Cluster at the sampling level, provided you have a small fraction of sampled clusters or a small fraction of sampled units within each cluster. This case is specifically important in panel data analysis, as we discuss in Section 11.4.4. Do not cluster in other cases.
4. **Sampling clustered, treatment clustered.** Cluster at the higher of the two levels to be conservative.

Let us go over two practical examples.

- **Case 1 (sampling cluster).** Some household or firm survey will (1) randomly select 50 out of 350 prefectures in China, and then (2) randomly select 100 households or firms in each sampled prefecture. This two stage design gives you a naturally stratified dataset. Just cluster at the prefecture level; in general, at the *first sampling stage* level.
- **Case 2 (treatment cluster).** STAR assigns treatment randomly at the class level. Then just cluster at the class level.

Choosing the Cluster Level

Ask two questions about the design, not about the residuals: How was the sample drawn? How was the treatment assigned? Cluster at the first sampling stage if sampling is clustered; cluster at the treatment assignment level if treatment is clustered; take the higher level if both; do not cluster if neither. Correlated residuals alone are not a reason to cluster, and unnecessary clustering is not harmless.

11.4.4 DID and Serial Correlation

One special case we must underscore is panel data analysis. When using panel data, we usually employ time variation for identification, as in the DID designs of Chapters 8 and 9. Think about how such a sample arises: you draw *people* (or provinces, or firms), not people in a specific year. Once individual i is in the panel, all of the periods of individual i come with them. This means you are drawing samples, and often assigning

treatment, clustered at the *individual* level, and the errors of the same individual are serially correlated across periods (Bertrand et al., 2004).

Thus, DID gives a natural clustering structure of the error, and case 3 of Section 11.4.3 applies with the individual as the sampling cluster.

The One Level Up Principle

In panel data, cluster at the individual, province, or city level, but **never** at the individual-year, province-year, or city-year level. Clustering at the unit by time level treats each period of the same unit as an independent draw, which is exactly the serial correlation mistake: the sampling unit is the individual with their whole time path, not the individual year cell. Cluster one level up from the unit time cell.

11.5 Comparing the Moulton and the AAIW Frameworks

Sections 11.3 and 11.4 approached the clustering problem from two directions. The framework of Moulton (1986), as presented in Angrist and Pischke (2009), and the framework of Abadie et al. (2023) are not two steps of one argument: they answer different questions under different thought experiments, and it is worth comparing them explicitly.

Consider first the two thought experiments. In the Moulton framework, randomness lives in the error term. The thought experiment is that nature redraws the shocks ν_g and η_{ig} from the error components model, the estimand is a parameter of the data generating process, in effect a feature of an infinite superpopulation, and the question is a diagnostic one: given correlated errors, how wrong is the conventional variance formula? Propositions 11.3.1 and 11.3.2 are the answer.

In the AAIW framework, potential outcomes are fixed attributes of a finite population; the only random objects are who is sampled, $\mathbf{R}$, and who is treated, $\mathbf{X}$. The survey redraws the sample, the policy redraws the treatment, the estimand is the average treatment effect of the population at hand, and the question is a constructive one: which variance matches this estimand, given how the data arose? Clustering enters if and only if one of the two processes operated in cluster blocks. Neither thought

experiment is wrong; they price different risks for different targets. The AAIW position is that the finite population causal estimand is the one applied papers actually claim: the estimand first discipline of Section 11.2, applied to clustering.

The bridge between the frameworks hides inside Proposition 11.3.2: the bias is governed by the *product* $\rho_x \rho_e$. Here, ρ_e belongs to nature; the researcher does not choose it. But ρ_x belongs to the *design*: class level assignment mechanically forces $\rho_x = 1$, individual randomization forces $\rho_x = 0$. The Moulton algebra therefore already implies that correlated errors alone do nothing. The folk rule, "residuals are correlated within groups, therefore cluster," silently sets $\rho_x = 1$ because the textbook examples have group level regressors; the first misconception of Section 11.4 refutes that folk rule, not the formula. AAIW promote ρ_x from an estimated parameter to a primitive the researcher *knows* from the design, and let cluster level heterogeneity in potential outcomes and effects play the role of ρ_e. The dictionary between the two languages is

$$\underbrace{\rho_x \, \rho_e > 0}_{\text{Moulton}} \quad \Longleftrightarrow \quad \underbrace{\text{design clustered and clusters explain outcome variation}}_{\text{AAIW}}.$$

The AAIW rule is the safe half of this condition: it never asks you to verify the heterogeneity part, which is a property of the estimand rather than of the design.

Where do the frameworks then agree? They agree exactly where the Moulton analysis was born: the sampled clusters are a small fraction of a large population of clusters and treatment operates at the cluster level. Writing q for the share of the population's clusters in the sample, the overstatement of the Liang and Zeger variance derived by Abadie et al. (2023) is proportional to q, so as $q \to 0$ the clustered standard error is exactly right. The AAIW framework nests the classical one as this limiting case. They depart on three questions the Moulton framework cannot answer: the *level* of clustering (an error components model can be posited at any level, so higher feels safer; AAIW pin the level to the design and show that clustering above it misprices, as in the firm and province example of Section 11.4); the *finite population* (in the census illustration of Abadie et al., 2023, the truth is 5.91, robust averages 1.90, clustered 44.86, and CCV/TSCB land near the truth); and *visibility* (clustered assignment shows in the data, clustered sampling does not, so the sampling reason to cluster must come from the survey documentation).

Moulton and AAIW in One Box

Moulton diagnoses a formula under a **parametric model**: nature redraws the error components, and conventional inference is off by the factor $1 + [\,V(n_g)/\bar{n} + \bar{n} - 1\,]\rho_x\rho_e$. AAIW construct the correct variance for the finite population causal estimand **from the design non-parametrically**: cluster if and only if sampling or assignment is clustered, at the level of the clustered process. The dictionary is $\rho_x \leftrightarrow$ the design and $\rho_e \leftrightarrow$ cluster level outcome heterogeneity. The frameworks coincide when few clusters are sampled and treatment is cluster assigned; they separate on the choice of level, on samples covering much of the population, and on what the data can reveal.

11.6 Conclusion

Today we discussed two nonstandard standard error issues: when the sample is large compared with the population, and when errors are not i.i.d. but clustered.

In the first issue, we claim that we need to consider both sampling-based and design-based uncertainty. Using traditional inference when the sample is a large share of a finite population produces standard errors that are too large and conservative; the overstatement is governed by treatment effect heterogeneity, and a finite population correction exists.

In the second case, we find that *not* adjusting for clustering generates standard errors that are too small when treatment or sampling is clustered. We can use the Liang and Zeger estimator to fix it, consistent as the number of groups grows. But clustering at a higher level is not always good: clustering comes from either clustered sampling or clustered treatment, so cluster at the first sampling stage or at the treatment assignment level, and do **not** cluster if you have a totally random sample with totally random treatment. In DID, cluster one level up to take care of the serial correlation.

The two halves of the chapter are one lesson: the standard error is a statement about a thought experiment, redrawing the sample, reassigning the treatment, or both, and you must know which thought experiment matches your design before you can defend the number in parentheses under your estimate.

Homework Problems

1. **Deriving the Intraclass Correlation Coefficient.**

   Consider the error structure of Section 11.3.2: $e_{ig} = \nu_g + \eta_{ig}$, where ν_g and η_{ig} are independent of each other, η_{ig} is independent across students, both components have mean zero, and homoskedasticity holds with variances σ_ν^2 and σ_η^2.

   (a) Compute the error variance $\sigma_e^2 = \text{var}(e_{ig})$.

   (b) Compute the covariance $E[e_{ig}e_{jg}]$ for two different students $i \neq j$ in the same class.

   (c) Combine (a) and (b) to derive equation (11.6): $\rho_e = \sigma_\nu^2/(\sigma_\nu^2 + \sigma_\eta^2)$.

   (d) Interpret ρ_e in one sentence, and state what error structures the two extreme values $\rho_e = 0$ and $\rho_e = 1$ correspond to.

2. **Deriving the Moulton Factor.**

   Consider the regression $y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}$ with G classes of equal size n, the regressor fixed within each class, and the error structure of Problem 1.

   (a) Show that the OLS estimator satisfies

   $$\hat{\beta}_1 - \beta_1 = \frac{\sum_{g=1}^{G}(x_g - \bar{x})\, n\, \bar{e}_g}{n\sum_{g=1}^{G}(x_g - \bar{x})^2}, \qquad \text{where } \bar{e}_g = \frac{1}{n}\sum_{i=1}^{n} e_{ig}.$$

   (b) Compute $\text{var}(\bar{e}_g)$ under the error structure of Problem 1, and express it in terms of σ_e^2, ρ_e, and n.

   (c) Derive the correct variance $V(\hat{\beta}_1)$, and compare it with the conventional OLS variance $V_c(\hat{\beta}_1) = \sigma_e^2 / \left(n\sum_g (x_g - \bar{x})^2\right)$ to conclude that

   $$\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)} = 1 + (n-1)\rho_e.$$

   (d) Evaluate the Moulton factor at $\rho_e = 1$ and explain, in terms of effective sample size, why the result makes sense.

3. **Choosing the Cluster Level.**

Answer in words, using the sampling and design framework of Abadie et al. (2023). For each scenario, state at which level (if any) you would cluster the standard errors, and justify your choice by identifying how the sample was drawn and how the treatment was assigned.

(a) A job training RCT randomly samples 5,000 unemployed workers from the national unemployment registry (millions of workers) and randomizes the training offer at the individual level. Workers who trained in the same center have correlated outcomes.

(b) The STAR experiment: students are randomly assigned to classes, and the class size treatment applies to the whole class.

(c) A firm survey first randomly selects 50 out of 350 prefectures, then randomly samples 100 firms within each selected prefecture; the regressor of interest is a firm level characteristic that is not assigned by any policy.

(d) A province-year panel of all 31 mainland Chinese provinces over 20 years, with a policy adopted by different provinces in different years (staggered DID). Explain both the level you would cluster at and why clustering at the province-year level is wrong.

(e) In scenario (d), a colleague suggests clustering at the region level (7 large regions containing the 31 provinces) "to be even more conservative." Using the effective sampling rate logic of Abadie et al. (2020), explain what is wrong with this advice.

Practice Example: The Moulton Problem in a Simulated Class Experiment

In this practice example we simulate a STAR style class level experiment, watch the conventional standard error understate the truth by exactly the Moulton factor, verify the fix with the Liang and Zeger estimator, and confirm by Monte Carlo that conventional confidence intervals undercover while clustered ones do not. We also verify the $\rho_x = 0$ lesson: with individual level treatment, no clustering is needed even though the errors remain clustered.

Data generating process. $G = 50$ classes, each with $n = 20$ students (1,000 obser-

vations). A binary class level treatment x_g is randomly assigned to exactly half of the classes. The outcome is

$$y_{ig} = 1 + 2\,x_g + \nu_g + \eta_{ig}, \qquad \nu_g \sim \mathcal{N}(0, 0.3), \quad \eta_{ig} \sim \mathcal{N}(0, 0.7),$$

so $\sigma_e^2 = 1$ and the intraclass correlation is $\rho_e = 0.3$ by equation (11.6). The predicted Moulton factor is $1 + (n-1)\rho_e = 1 + 19 \times 0.3 = 6.7$, so the correct standard error is about $\sqrt{6.7} \approx 2.59$ times the conventional one.

Analysis steps. The do-file (`src/chapter 11/ch11_cluster.do`) performs:

1. Simulate the dataset and save it; estimate ρ_e from the simulated components as a check.
2. Run OLS of y on x with conventional standard errors and with `vce(cluster class)`; compare the ratio of the two variance estimates with the predicted Moulton factor 6.7.
3. Run the group level regression (class means, WLS with class sizes) and compare with the Liang and Zeger standard error.
4. Monte Carlo with 500 replications: record the coverage rate of the nominal 95 percent confidence interval for $\beta_1 = 2$ under conventional and clustered standard errors. Conventional coverage should be far below 95 percent (about 55 percent at this Moulton factor); clustered coverage should be close to 95 percent.
5. Re run the Monte Carlo with the treatment randomized at the *individual* level (same clustered error structure): both conventional and clustered coverage should now be close to 95 percent, verifying that clustering of the errors alone does not require clustering the standard errors ($\rho_x = 0$).
6. Key lessons summary.

Key lessons.

- With group level treatment, the conventional standard error can understate the truth by a large factor, here $\sqrt{6.7} \approx 2.6$, and the miss is predicted exactly by the Moulton formula.
- The Liang and Zeger estimator repairs coverage without modeling the error structure, with consistency in the number of clusters.

- Clustered errors alone do not force clustered standard errors: with individual level assignment, $\rho_x = 0$ and conventional inference is fine, the first misconception of Abadie et al. (2023) in action.

The do-file is located at `src/chapter 11/ch11_cluster.do` and the simulated dataset at `data/chapter 11/`.

Chapter 12

Discrete Choice Model I

12.1 Introduction

In the previous lectures we focused on the reduced form approach: finding credible variation, comparing treated and untreated units, and interpreting a regression coefficient as a causal effect. In the following two chapters we give a very brief introduction to the *Discrete Choice Model* (DCM). The DCM considers problems where the outcome y is discrete: work or not work, take the bus or take a taxi, enter a market or stay out. Such outcomes are everywhere in economics, and they do not fit comfortably into the linear regression toolkit we have been using.

The DCM sits at the intersection of the reduced form and structural traditions, and it is an important method for both. You can learn and understand it in either framework:

- **Understood in a reduced form way**, the DCM is another kind of non-linear regression model. It is harder to interpret than OLS, but it fits better than the linear probability model when y is binary: predicted probabilities stay between zero and one, and the functional form respects the discreteness of the outcome.
- **Understood in a structural way**, the DCM opens a brand new world. Each parameter is a structural parameter of an underlying behavior model: a marginal utility, a cost coefficient, a taste. Because the model is built from utility maximization, it carries welfare implications: we can ask not only what agents choose, but how much they value their choices, and what a policy that changes the choice set would do to their welfare.

This chapter builds the model from the ground up: a motivating example, the general setting, what can and cannot be identified, and then the workhorse special case, the *Logit model*, together with its generalization, the *Nested Logit model*. A warning to keep in mind throughout: the Logit model is intrinsically structural, even when it is run casually as "just another regression."

12.2 Motivating Example: Female Labor Participation

Still remember the example in our first chapter? Consider a female labor participation problem. Individual i chooses labor supply to maximize utility over consumption and leisure,

$$\max_{l_i} U_i(c_i, 1 - l_i) + \epsilon_{il} \qquad \text{s.t.} \qquad c_i = w_i l_i, \tag{12.1}$$

where c_i is consumption, l_i is labor supply, ϵ_{il} is an unobserved taste shock attached to the labor choice, and w_i is the wage.

Assume that the labor choice is binary: work ($l_i = 1$) or not work ($l_i = 0$). Working delivers consumption w_i and no leisure; not working delivers no consumption and full leisure. The woman works if the utility of working exceeds the utility of not working,

$$U_i(w_i, 0) + \epsilon_{i1} \;\geq\; U_i(0, 1) + \epsilon_{i0}. \tag{12.2}$$

Given the wage w_i, this is a threshold rule in the taste shocks: define

$$\epsilon^*(w_i) = U_i(w_i, 0) - U_i(0, 1),$$

and the decision becomes

$$l_i = 1 \quad \text{if} \quad \epsilon_{i0} - \epsilon_{i1} < \epsilon^*(w_i). \tag{12.3}$$

The systematic part of the comparison, how much better working is in observables, sets a bar ϵ^*; the unobserved taste difference $\epsilon_{i0} - \epsilon_{i1}$ decides whether individual i gets over it.

Now let the taste difference $\epsilon_{i0} - \epsilon_{i1}$ have a CDF F in the population. The working

probability at wage w is

$$G(w) = \Pr(l = 1 \mid w) = \int_{-\infty}^{\epsilon^*(w)} dF = F\big(\epsilon^*(w)\big). \tag{12.4}$$

The observed participation rate is the CDF of the unobserved heterogeneity, evaluated at the observable utility gap.

Now, remind yourself of Chapter 1: what does the reduced form approach do with this problem? What does the structural approach do? What are the pros and cons of the two methods? The reduced form approach looks for credible variation in w and estimates how participation responds, without ever specifying U_i or F. The structural approach specifies the utility function and the shock distribution, estimates their parameters, and can then simulate counterfactual policies and compute welfare. This trade off, credibility of a single local effect versus the ability to extrapolate and evaluate welfare, is the running theme of the course.

This is a very typical example of a Discrete Choice Model. Today we give a brief introduction to the DCM and its most important example, the Logit model. One tip before we start: the Logit model is intrinsically structural. Everything in it, as we will see, is derived from a utility maximization problem.

12.3 The General Discrete Choice Model

12.3.1 Setting

The DCM describes decision makers' choices among discrete alternatives. A man chooses whether to smoke or not. A student chooses how to go to school: bus, taxi, or bike. A firm chooses whether to enter a local market: Walmart versus the local store. In each case, the choice set is a finite list, not a continuum.

This discreteness changes the mathematics of optimization. In a continuous and differentiable choice model, we optimize by taking first order conditions and finding an interior solution. Can we do the same thing for the DCM? **No.** There is no marginal adjustment between taking the bus and taking a taxi; the agent simply picks the best element of the list. The model must therefore be built on *comparisons* of utilities rather than derivatives of utilities.

Assume we have N decision makers, each choosing among a set of J alternatives $\{1, 2, \ldots, J\}$. Decision maker n gets utility U_{nj} from choosing alternative j, and the optimization is:

$$n \text{ chooses } i \quad \Longleftrightarrow \quad U_{ni} > U_{nj}, \ \ \forall j \neq i. \tag{12.5}$$

The researcher does not observe utility directly. What we see are the choice results, revealed preference, together with two kinds of observables: attributes of the alternatives faced by the agents, x_{nj} (the price of the bus ticket, the speed of the taxi), and personal characteristics of the agents, s_n (income, age). We collect the part of utility that can be explained by these observables into the *representative utility*

$$V_{nj} = V(x_{nj}, s_n),$$

and write the utility of choice j to agent n as

$$U_{nj} = V_{nj} + \epsilon_{nj}, \tag{12.6}$$

where ϵ_{nj} is the part of utility affected by factors the researcher does not observe. From the researcher's point of view ϵ_{nj} is a random variable, even though the agent knows it perfectly well.

12.3.2 The Choice Probability

Assume the vector of unobserved components $\epsilon_n' = [\epsilon_{n1}, \ldots, \epsilon_{nJ}]$ has a joint density $f(\epsilon_n)$ across the population. Then the probability that an agent with representative utilities $(V_{n1}, \ldots, V_{nJ})$ chooses alternative i is

$$\begin{aligned} P_{ni} &= P(U_{ni} > U_{nj}, \ \forall j \neq i) \\ &= P(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj}, \ \forall j \neq i) \\ &= P(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}, \ \forall j \neq i) \\ &= \int_{\epsilon} I(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}, \ \forall j \neq i)\, f(\epsilon_n)\, d\epsilon_n. \end{aligned} \tag{12.7}$$

Read the integral as a census: go through every possible realization of the unobserved tastes, keep the ones for which alternative i wins the utility comparison, and weight them by the density.

Different assumptions on the density $f(\epsilon_n)$ give different models, and expression (12.7) does not guarantee a closed-form choice probability in general:

- The Type I Extreme Value distribution gives the **Logit** model, with a closed-form probability.
- The Normal distribution gives the **Probit** model, with no closed-form probability.

Logit and Probit are therefore not competing frameworks; they are specific parametric instances of the same DCM.

12.4 Identification: What Can Utility Mean?

The identification of the DCM is important, and it relates to primitive properties of the utility function itself. Everything can be concluded in two statements:

1. **Only differences in utility matter.**
2. **The scale of utility is arbitrary.**

Why is this the case? Let us go back to the fundamental theory of utility.

12.4.1 Utility Represents Preference

Utility functions come from preferences. Assume we have a goods set X and a preference relation $\succsim$ defined on X, satisfying:

1. **Completeness:** for all $x, y \in X$, we have $x \succsim y$ or $y \succsim x$ (or both);
2. **Transitivity:** for all $x, y, z \in X$, if $x \succsim y$ and $y \succsim z$, then $x \succsim z$.

We call such a preference *rational*.

Definition 12.4.1 (Utility Representation, Definition 1.B.2 in Mas-Colell, Whinston, and Green, 1995)**.** A function $u : X \to \mathbb{R}$ is a utility function representing preference $\succsim$ if, for all $x, y \in X$,

$$x \succsim y \iff u(x) \geq u(y).$$

If there exists a utility function representing $\succsim$, then $\succsim$ is rational. A utility function assigns a numerical value to each element of X in accordance with the individual's preferences: utility is a *representation* of preference. And because preference is ordinal, utility is ordinal. If a rational preference can be represented by u, then it can also be

represented by any strictly increasing transformation of u: by $u+1$, by $u+k$, by $2u$, by ku, and so on. The numbers themselves carry no meaning; only the ranking they encode does. The two identification statements are exactly this ordinality, translated into the DCM.

12.4.2 Statement 1: Only Differences in Utility Matter

Let us use an example to reveal the two statements. Assume you can go to school either by bus (b) or by car (c). Let T_j be the speed of choice j and k_j the choice amenity:

$$U_c = \beta T_c + k_c + \epsilon_c,$$
$$U_b = \beta T_b + k_b + \epsilon_b.$$

The choice depends only on the difference:

$$U_c - U_b = \beta(T_c - T_b) + (k_c - k_b) + (\epsilon_c - \epsilon_b).$$

Only the difference $(k_c - k_b)$ can be identified, not k_c and k_b separately. The systems $\{u_j\}$ and $\{u_j+1\}$ are observationally equivalent: I do not care whether the comparison is $u_i - u_j$ or $(u_i+1)-(u_j+1)$. Thus you cannot give each alternative its own free constant. What to do in practice: *normalize the utility of one alternative to zero.* This is implicitly done for you whenever you run a logit or probit regression.

In addition, not all differences matter. Assume you include a personal characteristic Y_n (say, income) in the utility of each alternative:

$$U_{nc} = \beta T_c + \gamma Y_n + \delta Y_n T_c + \epsilon_{nc},$$
$$U_{nb} = \beta T_b + \gamma Y_n + \delta Y_n T_b + \epsilon_{nb},$$
$$U_{nb} - U_{nc} = \beta(T_b - T_c) + \delta Y_n(T_b - T_c) + (\epsilon_{nb} - \epsilon_{nc}).$$

The level term γY_n cancels out: δ is identified but γ is not. Differences in personal characteristics do not matter on their own, because we are comparing alternatives *for each person*, not comparing people with each other. A personal-level variable matters only if it interacts with a choice level variable, as $Y_n T_j$ does. The practical rule: do not add a personal-level variable without interacting it with a choice-level variable.

12.4.3 Statement 2: The Scale of Utility Is Arbitrary

Similarly, the systems $\{u_j\}$ and $\{2u_j\}$ are observationally equivalent: I do not care whether the comparison is $u_i - u_j$ or $2(u_i - u_j)$. Assume we have model 1,

$$\begin{aligned} U_{nc} &= \beta T_c + \gamma Y_n + \epsilon_{nc}, \\ U_{nb} &= \beta T_b + \gamma Y_n + \epsilon_{nb}, \\ U_{nb} - U_{nc} &= \beta(T_b - T_c) + (\epsilon_{nb} - \epsilon_{nc}), \end{aligned}$$

and model 2, which multiplies everything by two,

$$\begin{aligned} 2U_{nc} &= 2\beta T_c + 2\gamma Y_n + 2\epsilon_{nc}, \\ 2U_{nb} &= 2\beta T_b + 2\gamma Y_n + 2\epsilon_{nb}, \\ 2U_{nb} - 2U_{nc} &= 2\beta(T_b - T_c) + 2(\epsilon_{nb} - \epsilon_{nc}). \end{aligned}$$

Every choice made under model 1 is made identically under model 2, so the two are observationally equivalent, and the coefficients are only identified up to the common scale.

Thus we need to normalize the scale, and the standard way to do it is to *normalize the variance of the error term*. In Logit, this is automatically done: the Type I Extreme Value error has variance $\pi^2/6$. In Probit, this is automatically done: the standard Normal error has variance 1. This is worth remembering when comparing coefficients across models or across samples: the scale normalization is baked into the estimates.

12.5 The Logit Model

12.5.1 Setting

Following McFadden (1974), assume that each ϵ_{nj} is i.i.d. Type One Extreme Value (T1EV), with

$$\text{PDF: } f(\epsilon_{nj}) = e^{-\epsilon_{nj}} e^{-e^{-\epsilon_{nj}}}, \qquad \text{CDF: } F(\epsilon_{nj}) = e^{-e^{-\epsilon_{nj}}}.$$

Since the error terms are independent across alternatives, the joint CDF is the product

$$F(\epsilon_{n1}, \ldots, \epsilon_{nJ}) = \prod_{j=1}^{J} e^{-e^{-\epsilon_{nj}}}.$$

When the DCM is closed with this distributional assumption, we call it a *Logit model*.

12.5.2 Choice Probability

Let us derive the choice probability of the Logit model, starting from the general expression (12.7). It turns out that the integral can be solved in closed form.

Logit Choice Probability

Under i.i.d. T1EV errors, the (multinomial) choice probability is

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}. \tag{12.8}$$

Usually we normalize one of the choices, say j_0, to have a utility of zero, which gives

$$P_{ni} = \frac{e^{V_{ni}}}{1 + \sum_{j \neq j_0} e^{V_{nj}}}. \tag{12.9}$$

Deriving equation (12.8) is Homework Problem 1; the answer is in Chapter 3 of Train (2009).

In the binary choice case, the formula collapses to the familiar logistic function,

$$P_{n1} = \frac{e^{V_{n1}}}{1 + e^{V_{n1}}}. \tag{12.10}$$

The normalized choice is usually some baseline choice or an *outside option*. For instance, in an education choice model with choices {go to PKU, go to Fudan, go to SUFE, not go to school}, we normalize "not go to school" to have utility zero, and the other three utilities are measured relative to it.

What does the choice probability mean? The choice probability of i is the proportion of i's exponentiated choice value over the total exponentiated choice value of all alternatives. It is automatically compatible with the definition of a probability: $0 < P_{ni} < 1$

and $\sum_i P_{ni} = 1$. This is not like the linear probability model, which happily predicts probabilities below zero or above one.

The relation between the choice probability and the representative utility is sigmoid (S shaped), as Figure 12.1 shows for the binary case. The marginal effect of V_{ni} on P_{ni} increases first and then decreases: it is largest in the middle, where the agent is close to indifferent, and smallest in the tails, where the choice is nearly deterministic. Here is a question worth pausing on: if you fit this curve with a straight line, which part do you fit best? The middle, where the curve is approximately linear, and this observation will return when we compare Logit with the linear probability model in Section 12.7.

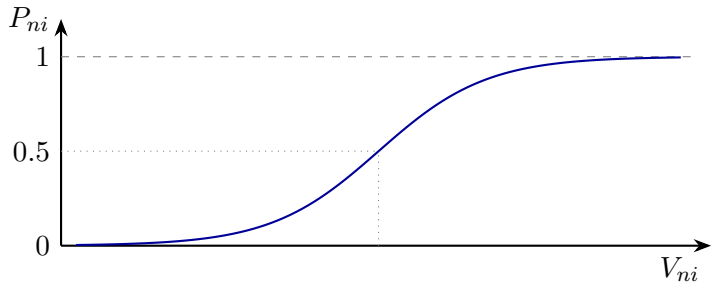


Figure 12.1: The sigmoid relation between representative utility and the Logit choice probability in the binary case. The slope is steepest at $P_{ni} = 0.5$ and flattens in both tails.

12.5.3 Independence from Irrelevant Alternatives

An important property of the Logit model is *Independence from Irrelevant Alternatives* (IIA). For any two alternatives i and k, the ratio of the Logit probabilities is

$$\frac{P_{ni}}{P_{nk}} = \frac{e^{V_{ni}} / \sum_j e^{V_{nj}}}{e^{V_{nk}} / \sum_j e^{V_{nj}}} = \frac{e^{V_{ni}}}{e^{V_{nk}}} = e^{V_{ni} - V_{nk}}. \tag{12.11}$$

The denominator, the only place where the other alternatives appear, cancels. The ratio has nothing to do with any other alternative: the probability ratio between any pair of choices depends only on their own choice values. Adding a new choice or deleting another choice will not change the ratio between i and k.

A manifestation of IIA is *proportionate shifting*: a change in an attribute z of choice j changes the probabilities of all other choices by the same proportion. With linear utility, the elasticity of the choice probability of i with respect to a change in z_{nj} of

choice j is

$$E_{iz_{nj}} = \frac{\partial P_{ni}}{\partial z_{nj}} \frac{z_{nj}}{P_{ni}} = -\beta_z \, z_{nj} \, P_{nj}, \qquad \forall i \neq j. \tag{12.12}$$

The expression is related only to alternative j; it is the same for every i. When the bus gets cheaper, the Logit model takes probability from taxis, bikes, and walking in equal proportion.

Is IIA a good property? Sometimes yes, sometimes no. It can save a great deal of computational resources when the number of choices is large, precisely because cross effects need not be tracked pair by pair. But it is also limiting, as the famous red bus and blue bus problem shows; we return to it in Section 12.6, where a more flexible model relaxes the property.

12.5.4 Derivatives and Marginal Effects

The derivative of the choice probability with respect to its own attribute is

$$\frac{\partial P_{ni}}{\partial z_{ni}} = \frac{\partial V_{ni}}{\partial z_{ni}} \, P_{ni}(1 - P_{ni}). \tag{12.13}$$

Deriving equation (12.13) is Homework Problem 2; the answer is again in Chapter 3 of Train (2009).

The Parameter Is Not the Marginal Effect

In the Logit model,

$$\frac{\partial P_{ni}}{\partial z_{ni}} \neq \frac{\partial V_{ni}}{\partial z_{ni}}.$$

Even if V is linear in z with coefficient $\beta = \partial V_{ni}/\partial z_{ni}$, you cannot interpret β as the marginal effect of z on the choice probability. The marginal effect is $\beta P_{ni}(1-P_{ni})$: it is non-linear, it differs across individuals, and it is largest when $P_{ni} = 1 - P_{ni} = 0.5$. Reporting Logit coefficients as if they were marginal effects is one of the most common mistakes in applied work.

12.5.5 Consumer Surplus

We are usually interested in the overall welfare of a consumer: what is the impact of some policy that changes some of the choices? In the Logit model, we have a closed-form solution for the expected utility that the agent obtains from the whole choice

set:

$$E(U_n) = E\Big[\max_j \big(V_{nj} + \epsilon_{nj}\big)\Big] = \ln\Big(\sum_{j=1}^{J} e^{V_{nj}}\Big) + C, \tag{12.14}$$

where C is a constant depending on the normalization. The expected utility is the *log sum* of the exponentiated values of all choices. Note what is being computed: the expectation of a maximum, taken before the taste shocks are realized, so this is the ex ante value of facing the menu. To convert utility into money, we can divide it by the marginal utility of income α_n:

$$E(CS_n) = \frac{1}{\alpha_n} E(U_n). \tag{12.15}$$

This is what we call "willingness to pay" in applied work.

The Two Closed Forms of Logit

There are two important closed-form formulas in the Logit model:

1. A closed-form **choice probability**: $P_{ni} = \dfrac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$;
2. A closed-form **expected (ex ante) utility** of the choice set: $E(U_n) = \ln\Big(\sum_{j=1}^{J} e^{V_{nj}}\Big) + C$.

They are very useful tricks in structural model research: the first turns utility parameters into predicted market shares, and the second turns an estimated model into a welfare calculator.

12.6 The Nested Logit Model

12.6.1 Motivating Example: Blue Bus versus Red Bus

In the previous section, we assumed that all alternatives are at the same level, with independent taste shocks. What if they have a hierarchical structure? As we have shown, Logit has the IIA property: given two options A and B, changes involving a third option do not change the relative probability of A and B. In some situations this property is not plausible.

Assume we have two choices, blue bus versus taxi, and the commuter is indifferent on

average:

$$P_{BB} = P_T = \frac{1}{2}.$$

One day, the bus company decides to introduce some buses with a new color, red. Now we have three choices: blue bus, red bus, taxi. A red bus is identical to a blue bus besides the color, so $P_{RB} = P_{BB}$. Due to IIA, the three probabilities must then be

$$P_{RB} = P_{BB} = P_T = \frac{1}{3}.$$

The intuitive answer is different: painting some buses red should not move anyone from taxis to buses, so we expect $P_{RB} = P_{BB} = 1/4$ and $P_T = 1/2$. Under Logit, you increase the probability of choosing the bus by basically doing nothing. The model fails here because the two buses' unobserved components are nearly perfectly correlated, while Logit insists all shocks are independent.

12.6.2 Setting

To solve the blue and red bus issue, we introduce an extension of the Logit model: the *Nested Logit* model, which allows for correlations among some of the options. Utility keeps the additive form

$$U_{nj} = V_{nj} + \epsilon_{nj}, \tag{12.16}$$

but now the vector $\epsilon = (\epsilon_{n1}, \ldots, \epsilon_{nJ})$ is jointly distributed as a *Generalized Extreme Value* (GEV) distribution (McFadden, 1978).

Let the choice set be partitioned into K subsets $B_1, \ldots, B_K$, called *nests*. The CDF of ϵ is

$$F(\epsilon) = \exp\left(- \sum_{k=1}^{K} \left(\sum_{j \in B_k} e^{-\epsilon_{nj}/\lambda_k} \right)^{\lambda_k} \right). \tag{12.17}$$

This distribution has the following properties:

- The marginal distribution of each ϵ_{nj} is univariate T1EV;
- Any two options within the same nest have correlated ϵ;
- Any two options in different nests have uncorrelated ϵ;
- λ_k is a measure of the degree of independence within nest k: a higher λ_k means less correlation of choices within the same nest.

Homework Problem 3 asks: what does it mean when $\lambda_k = 1$ for all k? What is the model then, and why.

Figure 12.2 shows the classic example: a commuter first chooses between the Auto nest and the Transit nest, and then an alternative within the nest.

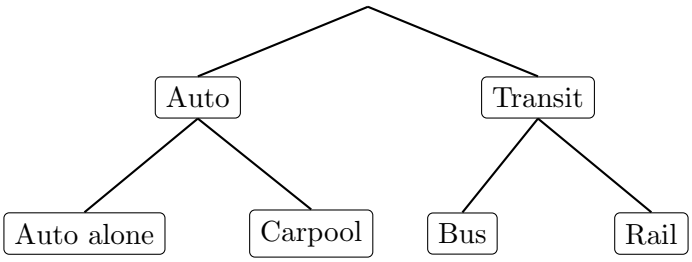


Figure 12.2: Tree diagram for mode choice with two nests: Auto = (Auto alone, Carpool) and Transit = (Bus, Rail). Adapted from Figure 4.1 in Train (2009).

12.6.3 Choice Probability

It can be shown that the choice probability of the Nested Logit model, for an alternative i belonging to nest B_k, is

$$P_{ni} = \frac{e^{V_{ni}/\lambda_k} \left(\sum_{j \in B_k} e^{V_{nj}/\lambda_k} \right)^{\lambda_k - 1}}{\sum_{l=1}^{K} \left(\sum_{j \in B_l} e^{V_{nj}/\lambda_l} \right)^{\lambda_l}}. \tag{12.18}$$

Note the structure. The term $\left(\sum_{j \in B_k} e^{V_{nj}/\lambda_k} \right)^{\lambda_k - 1}$ in the numerator collects all the choices in the same nest as i: how attractive alternative i is depends not only on its own value but on the value of its whole nest. As a consistency check, when $\lambda_k = 1$ for every nest, the inner sums enter with exponent zero in the numerator and exponent one in the denominator, and (12.18) collapses exactly to the plain Logit probability (12.8).

12.6.4 Independence from Irrelevant Nests

Given two alternatives $i \in B_k$ and $m \in B_l$, the probability ratio is

$$\frac{P_{ni}}{P_{nm}} = \frac{e^{V_{ni}/\lambda_k}\left(\sum_{j\in B_k} e^{V_{nj}/\lambda_k}\right)^{\lambda_k-1}}{e^{V_{nm}/\lambda_l}\left(\sum_{j\in B_l} e^{V_{nj}/\lambda_l}\right)^{\lambda_l-1}}, \tag{12.19}$$

because the common denominator of (12.18) cancels. Two cases follow:

- If $k = l$, so the two alternatives are in the *same nest*, the nest sums also cancel and we are back to IIA:
$$\frac{P_{ni}}{P_{nm}} = \frac{e^{V_{ni}/\lambda_k}}{e^{V_{nm}/\lambda_k}}.$$
- If $k \neq l$, so the two alternatives are in *different nests*, we do not have IIA: the relative probability of i and m depends on the other choices in their own nests B_k and B_l. But it still does not depend on choices in any *other* nest.

We call this property *Independence from Irrelevant Nests* (IIN). It is exactly the right amount of flexibility for the bus example: put the two buses in one nest, and adding the red bus reshuffles probability within the bus nest while leaving the bus versus taxi ratio governed by the nest values, not by the count of bus colors.

12.7 Logit or LPM?

An important practical question is: when should you use Logit, and when should you use the linear probability model (LPM), which is just OLS with a binary outcome? Let us first list the pros and cons.

For Logit, which is a non-linear fit with a functional form assumption:

- Coefficients are "structural" and primitive: they live in utility, production, and other economic objects;
- But coefficients are neither marginal effects nor weighted treatment effects, and translating them into effects takes extra work;
- It is computationally intensive: especially maximum likelihood with high dimensional fixed effects (many dummies), which can be slow or unstable.

For LPM, which is a linear fit and more of an approximation:

- Coefficients are marginal effects, very easy to interpret;
- But it will predict probabilities above one or below zero for extreme covariate values;
- It is computationally simple: it is just an OLS regression.

Choosing between Logit and LPM

Here are some personal views.

- If you care about the primitive parameter (a utility weight, a structural elasticity), use **Logit**.
- If you are interested in extrapolating your prediction, that is, predicting y at values of x with few samples nearby, use **Logit**: the functional form disciplines the extrapolation.
- If your x is distributed fairly uniformly over its range but you want to predict y at very small or very large x, use **Logit**: the linear fit is anchored to the middle of the sigmoid and misses the tails.
- Otherwise, you can choose **LPM**: in the middle of the distribution, where the sigmoid is approximately linear, the two models agree and the LPM's interpretability wins.

12.8 Conclusion

Here are some main takeaways of this chapter. First, DCM/Logit is intrinsically a structural approach: its parameters have structural meaning as utility weights, and the model is derived from utility maximization. Second, Logit is a special kind of DCM, the one obtained when the errors are i.i.d. T1EV distributed; Probit is the DCM with Normal errors. Third, Logit is convenient because it has a closed-form choice probability and a closed form expected utility, the log sum formula, which together make it a complete toolkit for prediction and welfare analysis.

Fourth, Logit has the IIA property: the relative probability of two choices is not affected by a third one. This is sometimes a convenience and sometimes a defect, as the red bus and blue bus example shows. Fifth, the interpretation of Logit, and of non-linear models in general, is not as straightforward as the linear probability model:

the coefficient is not the marginal effect. Sixth, Nested Logit is a more general model: with GEV errors, choices within the same nest have correlated shocks. IIA survives within a nest, but across nests it is replaced by IIN, independence from irrelevant nests.

We will further discuss the endogeneity issue in the DCM in the next lecture: what happens when the observed attributes are correlated with the unobserved tastes, and what the discrete choice analogue of instrumental variables looks like.

Homework Problems

1. **Deriving the Logit Choice Probability.**

   Assume that the unobserved utility components ϵ_{nj} are i.i.d. Type One Extreme Value across alternatives, with CDF $F(\epsilon) = e^{-e^{-\epsilon}}$ and PDF $f(\epsilon) = e^{-\epsilon}e^{-e^{-\epsilon}}$. Starting from the general choice probability

   $$P_{ni} = \int_{\epsilon} I(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj},\ \forall j \neq i)\, f(\epsilon_n)\, d\epsilon_n,$$

   derive the closed-form Logit choice probability

   $$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}.$$

   (Hint: condition on ϵ_{ni}, use independence to write the inner probability as a product of T1EV CDFs, and integrate over ϵ_{ni}. The answer is in Chapter 3 of Train, 2009.)

2. **Deriving the Own Derivative of the Choice Probability.**

   Starting from the Logit choice probability $P_{ni} = e^{V_{ni}} / \sum_j e^{V_{nj}}$, derive the derivative formula

   $$\frac{\partial P_{ni}}{\partial z_{ni}} = \frac{\partial V_{ni}}{\partial z_{ni}} P_{ni}\,(1 - P_{ni}),$$

   where z_{ni} is an attribute entering V_{ni}. Explain in one or two sentences why the marginal effect is largest at $P_{ni} = 0.5$ and what this implies for interpreting Logit coefficients. (The answer is in Chapter 3 of Train, 2009.)

3. **Nested Logit with $\lambda_k = 1$.**

Consider the Nested Logit model with GEV distribution

$$F(\epsilon) = \exp\left(-\sum_{k=1}^{K} \Big(\sum_{j \in B_k} e^{-\epsilon_{nj}/\lambda_k} \Big)^{\lambda_k} \right).$$

What does it mean when you have $\lambda_k = 1$ for all k? What is the model now? Why? Verify your answer both from the CDF and from the choice probability formula.

Practice Example: Logit, Marginal Effects, and the IIA Property

In this practice example we simulate discrete choices from a true utility maximization model, estimate Logit models on the simulated data, and verify the four central lessons of this chapter: the coefficient is not the marginal effect, the LPM and Logit agree in the middle of the distribution but disagree in the tails, and the Logit choice probabilities satisfy IIA.

Data generating process. Two designs.

1. *Binary participation.* $N = 5{,}000$ women draw a log wage $\ln w_i \sim \mathcal{N}(3, 0.25)$. Utility of working is $V_{i1} = -4 + 1.5 \ln w_i$ relative to the outside option $V_{i0} = 0$, and each option carries an i.i.d. T1EV shock, generated by the inverse transform $\epsilon = -\ln(-\ln(u))$ with $u \sim \text{Uniform}(0, 1)$. The woman works if $V_{i1} + \epsilon_{i1} > \epsilon_{i0}$. The implied true participation probability is the logistic function of V_{i1}.

2. *Transport mode choice.* $N = 5{,}000$ commuters choose among bus, taxi, and bike. Each mode j has a cost c_{ij} (in yuan) and a time t_{ij} (in hours), drawn from mode specific distributions:

$$\begin{array}{lll} \text{bus:} & c_{ij} \sim \text{Uniform}(1, 3), & t_{ij} \sim \text{Uniform}(0.5, 1.5); \\ \text{taxi:} & c_{ij} \sim \text{Uniform}(8, 15), & t_{ij} \sim \text{Uniform}(0.2, 0.6); \\ \text{bike:} & c_{ij} = 0, & t_{ij} \sim \text{Uniform}(0.8, 2.0). \end{array}$$

The representative utility is $V_{ij} = -0.8\, c_{ij} - 1.2\, t_{ij} + k_j$, with alternative constants $k_{bus} = 1$ and $k_{taxi} = 8$ and bike as the normalized base ($k_{bike} = 0$), plus i.i.d. T1EV shocks on each mode.

Analysis steps. The do-file (`src/chapter 12/ch12_logit.do`) performs:

1. Simulate the binary design; estimate a Logit by maximum likelihood and verify the coefficients recover the truth $(-4, 1.5)$.
2. Compare three objects: the raw Logit coefficient, the average marginal effect from `margins, dydx()`, and the LPM coefficient from OLS. Confirm the LPM coefficient approximates the average marginal effect, not the Logit coefficient.
3. Plot predicted probabilities from Logit and LPM against $\ln w$; confirm the LPM prediction leaves $[0, 1]$ in the tails while the Logit prediction does not.
4. Simulate the mode choice design; estimate a conditional logit with alternative specific attributes (`asclogit`); verify the cost and time coefficients recover the truth.
5. Verify IIA in the estimated model: compute $\hat{P}_{bus}/\hat{P}_{taxi}$ for each commuter, then re estimate after deleting the bike alternative, and confirm the ratio is identical up to estimation noise (it is exactly identical at the true parameters).
6. Compute the log sum welfare measure before and after a bus fare increase, converting to money with the cost coefficient as the marginal utility of income; report the average compensating variation.

Key lessons.

- The Logit coefficient is a utility parameter; the marginal effect on the probability is $\beta P(1 - P)$ and must be computed, not read off the coefficient.
- LPM and Logit agree where the sigmoid is flat in the middle; they part in the tails, where the LPM leaves the unit interval.
- The estimated Logit reproduces IIA up to estimation noise: probability ratios between two modes are exactly identical at the true parameters, and move only by estimation noise when a third mode is removed and the model is re estimated.
- The log sum formula turns the estimated utility model into a welfare calculator.

The do-file is located at `src/chapter 12/ch12_logit.do` and the simulated datasets at `data/chapter 12/`.

Chapter 13

Discrete Choice Model II

13.1 Introduction

In the previous chapter we built the Discrete Choice Model from utility maximization and derived the Logit model, with its closed-form choice probability

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}. \tag{13.1}$$

Two practical questions remain, and they are the subject of this chapter.

First, *how do we actually estimate the model?* The parameters live inside a non-linear probability, so ordinary least squares no longer applies directly. We introduce Maximum Likelihood Estimation (MLE) and, because the likelihood rarely has a closed-form maximizer, a brief tour of the numerical optimization methods that econometric software uses under the hood.

Second, *what if a regressor is endogenous?* Assume a car purchase problem for a consumer. The car price is correlated with unobserved product quality, and the familiar 2SLS recipe from the linear world cannot be transplanted into a non-linear model. We introduce three ways to use instruments in a DCM: the BLP approach of Berry, Levinsohn, and Pakes (1995), which converts the non-linear endogeneity problem into a linear one; the control function approach; and IV-Probit. Along the way we will meet two tools with life far beyond this chapter: the contraction mapping and the Banach fixed point theorem.

13.2 Maximum Likelihood Estimation

13.2.1 A Naive Approach: Log-linearization

We have introduced the Logit model; now we consider how to estimate it. There are several ways. A simple and naive way is to log-linearize the choice probability (13.1):

$$\ln P_{ni} = V_{ni} - \underbrace{\ln\Big(\sum_j e^{V_{nj}}\Big)}_{FE_n}, \tag{13.2}$$

where the log sum term is common to all alternatives faced by individual n. When V_{ni} is linear in parameters, we can therefore run an OLS regression of log choice shares on the attributes, with a fixed effect at the n level absorbing the log sum.

This trick is widely used in the trade literature, especially in estimating the gravity equation, and it is very easy to implement. However, it is **not** the main method people usually use, especially in labor economics, for four reasons:

1. In data, when an observed share is zero, $P_{ni} = 0$, the log value is undefined;
2. Granular settings create problems in spatial economics: with many locations and few migrants per cell, observed shares are noisy or zero (Dingel and Tintelnot, 2020);
3. Anything absorbed by the fixed effect is lost: parameters on n level variables in V_{ni} cannot be estimated;
4. The OLS uses only part of the information in the model, so it is not efficient.

13.2.2 The Likelihood Function

The best method is MLE. Let y_{ni} indicate whether choice i is chosen in the data by individual n. The likelihood of observing the whole pattern of choices, as a function of the parameter vector θ, is

$$L(\theta) = \prod_{n=1}^{N} \prod_i \big(P_{ni}(\theta)\big)^{y_{ni}}, \tag{13.3}$$

and taking logs gives the log likelihood

$$LL(\theta) = \sum_{n=1}^{N} \sum_{i} y_{ni} \ln P_{ni}(\theta). \tag{13.4}$$

The MLE estimator is

$$\hat{\theta}^{MLE} = \arg\max_{\theta} LL(\theta). \tag{13.5}$$

The interpretation is direct: we choose θ to maximize the probability of observing exactly the choice data y_{ni} that we did observe. Every observation contributes its full predicted probability, so nothing is thrown away, which is the efficiency advantage over the log-linearized OLS.

In OLS, optimization is simple: the first order condition is linear and we have a closed-form solution. But very often there is no closed-form solution in MLE: the log likelihood of a Logit model is a smooth non-linear function of θ, and its maximizer must be found numerically. This brings us to numerical non-linear optimization.

13.3 Numerical Optimization

13.3.1 Guess, Update, Iterate

The idea of numerical non-linear optimization is: guess, update, iterate.

- **Step 1:** make an initial guess θ_0;
- **Step 2:** use some updating rule to move θ_0 to θ_1, searching toward the optimum;
- **Step 3:** keep updating $(\theta_2, \theta_3, \ldots, \theta_t)$ until the sequence converges.

The question is: how to search and update? Here I give a very brief introduction to the gradient-based numerical methods used in economics. The answer, in one line, is: *follow the derivative.* Figure 13.1 shows the picture to keep in mind: the log likelihood is a hill, the current guess θ_t sits somewhere on the slope, and we want to climb to the top $\hat{\theta}$.

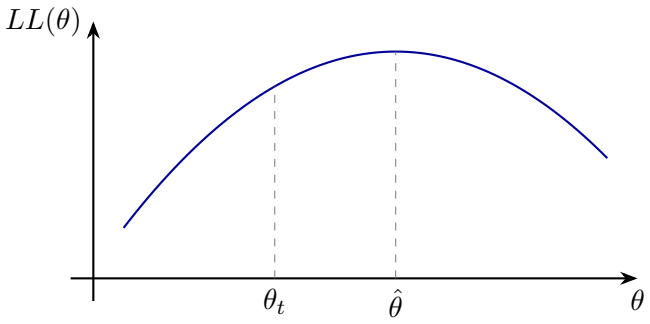


Figure 13.1: The log likelihood as a hill: the maximizer $\hat{\theta}$ sits at the top, and a numerical method climbs toward it from the current guess θ_t. Adapted from Figures 8.1 to 8.3 in Train (2009).

13.3.2 Gradient, Hessian, and the Updating Rule

Define the $K \times 1$ *gradient* at iteration t as the vector of first order derivatives of the log likelihood, evaluated at the current guess:

$$g_t = \left(\frac{\partial LL(\theta)}{\partial \theta} \right)_{\theta_t}. \tag{13.6}$$

Define the $K \times K$ *Hessian* at t as the matrix of second order derivatives, including all cross derivatives:

$$H_t = \left(\frac{\partial g(\theta)}{\partial \theta'} \right)_{\theta_t} = \left(\frac{\partial^2 LL(\theta)}{\partial \theta \, \partial \theta'} \right)_{\theta_t}. \tag{13.7}$$

All gradient-based optimization methods share the same form of updating rule:

$$\theta_{t+1} = \theta_t + \lambda \, M \, g_t, \tag{13.8}$$

where the gradient g_t controls the updating *direction*, and the scalar step size λ together with the $K \times K$ matrix M control the updating *speed*. Different choices of M define different methods.

13.3.3 Newton–Raphson

The most famous method is Newton–Raphson (NR), which uses the negative inverse Hessian as the speed matrix:

$$\theta_{t+1} = \theta_t + \lambda \left(-H_t\right)^{-1} g_t. \tag{13.9}$$

This choice is very intuitive, and Figure 13.2 illustrates both halves of the intuition:

- **The gradient tells us the direction.** A positive slope means the top is to the right, so increase θ_{t+1}; a negative slope means we have passed the top, so decrease θ_{t+1}.
- **The Hessian is the curvature of the function.** Where the function is sharply curved, the slope changes quickly, so a big jump would overshoot: be conservative. Where the function is flat, the slope changes slowly, so take a big step. Dividing the gradient by the curvature does exactly this: step size is inversely related to curvature.

The scalar λ helps us adjust the step in case the full NR step is too large and updates past the maximum. We could call it a damping parameter.

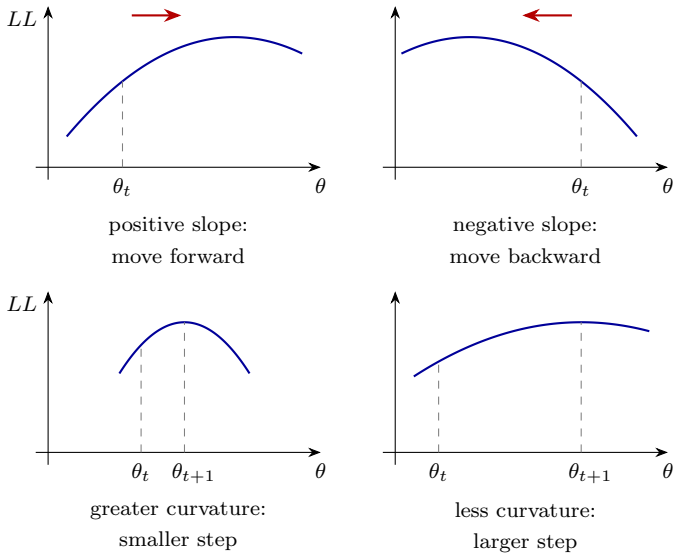


Figure 13.2: The two halves of the Newton–Raphson intuition. Top row: the sign of the gradient sets the direction of the step. Bottom row: the curvature (Hessian) sets its size, with sharper curvature calling for a smaller step. Adapted from Figures 8.1 to 8.3 in Train (2009).

There are two drawbacks of the NR method:

1. The calculation of the Hessian is computationally intensive: $K(K+1)/2$ second derivatives at every iteration;
2. There is no guarantee that the step follows the direction of the gradient if LL is not globally concave: away from the maximum, the Hessian is not necessarily negative definite, and $(-H_t)^{-1}g_t$ can point the wrong way.

There are other methods using different speed matrices M to overcome these issues, such as replacing the Hessian with guaranteed positive definite approximations built from gradients. For more details, please refer to Chapter 8 of Train (2009).

Optimization in Practice

When doing MLE, you can usually rely on well tested optimization packages: `Optim` in Julia, `fminsearch` in Matlab, or the `ml` machinery inside Stata. There is no need to code the updating rules yourself; what matters is understanding what the package is doing, so that when it fails to converge you know whether to blame the starting values, the scaling of the variables, or the shape of your likelihood.

13.4 Endogeneity in the DCM

So far we have assumed that the unobserved errors are independent of the explanatory variables. But this cannot always be the case. Consider consumers buying cars, the setting of Berry, Levinsohn, and Pakes (1995). The effect of the car price on the purchase probability is hard to identify, because there can be unobserved car traits correlated with prices: smoothness of the ride, brand reputation, interior quality. Expensive cars are expensive partly *because* they are good in ways the econometrician does not observe. This is the endogeneity issue in the DCM.

Linear IV Does Not Transplant

IV estimation as we know it from the linear world cannot be directly used in a non-linear model, even if we have a valid instrument. There is no residual to project out, no first stage to substitute in: the endogenous variable sits inside a non-linear choice probability, and mechanically running 2SLS on a Logit model produces neither consistent estimates nor interpretable ones.

How, then, can we use an IV in a non-linear model such as the DCM? We introduce three methods: (1) BLP; (2) the control function; (3) IV-Probit.

13.4.1 Method 1: BLP

13.4.1.1 Setup

The first method was introduced by Berry, Levinsohn, and Pakes (1995), and everyone calls it BLP. The idea is to transform the endogeneity issue in a non-linear model into a linear one, and then use the well developed linear IV method to solve it.

Assume we have the car buying problem. There are M markets (cities, years) with J_m options (brands) in each market. The utility for consumer n in market m choosing brand j is

$$U_{njm} = V(p_{jm}, x_{jm}, s_n, \beta_n) + \xi_{jm} + \epsilon_{njm}, \tag{13.10}$$

where p_{jm} is the price, x_{jm} are observed product attributes, s_n are personal attributes, ξ_{jm} are *unobserved product attributes*, and ϵ_{njm} is an i.i.d. T1EV shock. In the simple case V is linear. The endogeneity problem is that price is correlated with the unobserved brand attributes: $\xi_{jm} \not\perp p_{jm}$.

The important feature of BLP is that the endogeneity comes from the product-market level ξ_{jm}: the unobserved quality is a property of brand j in market m, common to all consumers there, not a property of the individual consumer.

13.4.1.2 The Two-step Idea: Pack, Estimate, Unpack

BLP employs a two-step approach: first, add a product-market level fixed effect that absorbs ξ_{jm}, and estimate the choice model with these fixed effects; second, open the box of the fixed effects and estimate the remaining parameters by linear IV. The key point is that the endogeneity happens *only* at the product-market level, so once it is packed into a fixed effect, the rest of the model is clean.

Decompose the observed utility value into a part that varies only over product-markets and a part that also varies over consumers:

$$V_{njm} = \underbrace{\bar{V}(p_{jm}, x_{jm}, \bar{\beta})}_{\text{varies only over product-market}} + \underbrace{\tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)}_{\text{varies also over consumer}}.$$

Then the utility is

$$U_{njm} = \left[\bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm}\right] + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \epsilon_{njm},$$

where the bracketed term is a product-market level fixed effect. Define it as

$$\delta_{jm} = \bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm}, \tag{13.11}$$

so that

$$U_{njm} = \delta_{jm} + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \epsilon_{njm}. \tag{13.12}$$

Equation (13.12) does not entail any endogeneity: everything correlated with ξ_{jm} is inside δ_{jm}, which is treated as a parameter to estimate. The procedure is:

- **Step 1:** run a Logit model with product-market level fixed effects to estimate the consumer-level parameters $\tilde{\beta}$ (and recover the estimated fixed effects $\hat{\delta}_{jm}$);
- **Step 2:** take the estimated $\hat{\delta}_{jm}$, and run a linear IV regression of equation (13.11), instrumenting the price, to estimate $\bar{\beta}$.

The Essence of BLP

We cannot run an IV regression directly in a DCM. So: first *pack* all terms at the level where the endogeneity happens into fixed effects; then *estimate* the DCM with these fixed effects, which is a clean non-linear problem; finally *unpack* the estimated fixed effects and run a linear IV regression on them. The non-linear IV problem has been transformed into a linear IV problem. BLP tells you how to use an IV in a DCM, but only in a specific model structure: the endogeneity must live at a coarser level than the individual choice.

13.4.1.3 High Dimensional Fixed Effects and the Berry Contraction

Step 1 raises its own problem. Sometimes we have many product-market combinations, so the dimension of the δ_{jm} vector is very high. Feeding tens of thousands of fixed effects into a traditional non-linear optimization algorithm inside the MLE is slow or even impossible.

Here we use the *Berry contraction* (Berry, 1994): a contraction mapping method to numerically compute the fixed effects, and a standard trick for solving high dimensional

fixed effects in general. The procedure rests on one observation: the fixed effects determine the predicted market share of each product in each market, so in each iteration we can push the predicted shares toward the actual shares in the data.

Let S_{jm} be the real market share of product j in market m in the data (e.g. the share of BYD in Shanghai). Define the predicted share from the model as

$$\hat{S}_{jm} = \sum_n \hat{P}_{njm}/N_m,$$

where $\sum_n \hat{P}_{njm}$ is the predicted total sales of product j in market m and N_m is the total sales in market m. Denote by δ the vector collecting δ_{jm} for all j, m. The algorithm for Step 1 is:

1. Take an initial guess of the consumer-level parameters $\tilde{\beta}_n$ in $\tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)$;
2. Take an initial guess of δ;
3. For each guess $\tilde{V}^t$ and δ^t, calculate the choice value for each consumer, product, and market: $\hat{U}^t_{njm} = \delta^t_{jm} + \tilde{V}^t(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)$;
4. Given the choice values, calculate the predicted choice probabilities and shares:

$$\hat{P}^t_{njm} = \frac{\exp\left(\hat{U}^t_{njm}\right)}{\sum_{j'} \exp\left(\hat{U}^t_{nj'm}\right)}, \qquad \hat{S}^t_{jm} = \sum_n \hat{P}^t_{njm}/N_m;$$

5. Given the predicted shares, update the fixed effects:

$$\delta^{t+1}_{jm} = \delta^t_{jm} + \ln\left(\frac{S_{jm}}{\hat{S}^t_{jm}(\delta^t)}\right); \tag{13.13}$$

6. Iterate steps 3 to 5 until the fixed effects converge, for the current guess of $\tilde{\beta}_n$;
7. Use a traditional non-linear optimization method to iterate over steps 1 to 6 until the optimum for $\tilde{\beta}_n$ is found.

The structure is a nested loop. In the *outer loop* (steps 1 to 6 as a whole), we estimate $\tilde{\beta}_n$ with a traditional MLE algorithm. In the *inner loop* (steps 3 to 5), for each candidate $\tilde{\beta}_n$, we iterate the fixed effects until the predicted shares match the observed shares. The update (13.13) has a transparent logic: if the model underpredicts a product's

share, $\hat{S}_{jm}^t < S_{jm}$, the log ratio is positive and the product's fixed effect is revised upward, making it more attractive in the next round.

13.4.1.4 Why It Works: Contraction Mappings and the Fixed Point

Why does this iteration converge, and why does it not depend on the initial guess? Because the Berry update is a *contraction mapping*.

Definition 13.4.1 (Contraction Mapping)**.** Let (X, d) be a metric space. A map $T : X \to X$ is called a contraction mapping on X if there exists $q \in [0, 1)$ such that

$$d\big(T(x), T(y)\big) \leq q\, d(x, y) \qquad \text{for all } x, y \in X.$$

A contraction mapping is a function that squeezes points closer together: apply T to any two points, and their distance shrinks by at least the factor q.

Theorem 13.4.1 (Banach Fixed Point Theorem)**.** *Let (X, d) be a non-empty complete metric space with a contraction mapping $T : X \to X$. Then T admits a unique fixed point x^* in X, that is, $T(x^*) = x^*$. Furthermore, x^* can be found as follows: start with an arbitrary element $x_0 \in X$ and define the sequence $x_t = T(x_{t-1})$ for $t \geq 1$; then* $\lim_{t\to\infty} x_t = x^*$.

The existence of a contraction mapping implies a unique fixed point, and the fixed point can be found by iterating the map from *any* starting point: x_0, $T(x_0)$, $T(T(x_0))$, $T(T(T(x_0)))$, and so on.

What does this mean in the BLP algorithm? As long as the updating iteration (13.13) is a contraction mapping, which Berry, Levinsohn, and Pakes (1995) prove for this model, we will converge to the same δ no matter what our initial guess δ^0 was. The terminal point is the fixed point $\delta = T(\delta)$, and at that point the predicted shares equal the shares in the data.

This idea is used far beyond BLP. Whenever we need to compute the equilibrium of a complicated model, a standard strategy is to write the equilibrium as the fixed point of a mapping, verify the mapping contracts, and iterate.

One caveat before moving on: BLP is not always feasible. The method leans on the specific error structure, with the endogenous unobservable living at the product-market level, and the estimation algorithm is genuinely involved. I highly recommend reading

the BLP chapter of Train (2009), or better, Berry, Levinsohn, and Pakes (1995) itself.

13.4.2 Method 2: The Control Function

The second important non-linear IV approach is the Control Function (CF). The utility of consumer n buying product j is

$$U_{nj} = V(x_{nj}, w_{nj}, \beta_n) + \epsilon_{nj},$$

where x_{nj} is endogenous, $x_{nj} \not\perp \epsilon_{nj}$, and w_{nj} are exogenous attributes. We assume there is an instrument z_{nj}, related to x_{nj} through a first stage:

$$x_{nj} = W(z_{nj}, \gamma) + \mu_{nj}. \tag{13.14}$$

Assume $\epsilon_{nj}, \mu_{nj} \perp z_{nj}$, but $\epsilon_{nj} \not\perp \mu_{nj}$. The latter is exactly what makes x_{nj} endogenous: x is correlated with the utility error only through the first stage error μ.

Now do a CEF decomposition of ϵ_{nj} given μ_{nj}:

$$\epsilon_{nj} = \underbrace{E(\epsilon_{nj} \mid \mu_{nj})}_{CF(\mu_{nj},\lambda)} + \tilde{\epsilon}_{nj}.$$

By construction, the CEF residual satisfies $\tilde{\epsilon}_{nj} \perp \mu_{nj}$, and therefore $\tilde{\epsilon}_{nj} \perp x_{nj}$: since x is correlated with ϵ only through μ, removing the part of ϵ explained by μ removes all of the correlation. We call $CF(\mu_{nj}, \lambda)$ a *control function*, where λ is some parameter vector. After controlling for $CF(\mu_{nj}, \lambda)$, x is no longer correlated with the remaining error.

The utility function becomes

$$U_{nj} = V(x_{nj}, w_{nj}, \beta_n) + CF(\mu_{nj}, \lambda) + \tilde{\epsilon}_{nj}, \tag{13.15}$$

and the estimation is a two-step procedure:

- **Step 1:** estimate the first stage equation (13.14) and obtain the first stage residuals $\hat{\mu}_{nj}$;
- **Step 2:** plug $\hat{\mu}_{nj}$ into the control function in (13.15), and estimate the model as a simple Logit with the control function as an extra regressor.

In Step 2 we need to assume a functional form for CF. Usually we choose a flexible non-parametric form, for example high order polynomials in $\hat{\mu}_{nj}$.

The logic of the CF approach, in three lines: the instrument z is not correlated with the error ϵ; the endogenous variable x therefore correlates with ϵ only through the first stage error μ; so by controlling for the part of ϵ that comoves with μ, we eliminate the correlation between x and the remaining error. The CF is a pretty general method, applicable well beyond the Logit model, but it requires you to take a stand on the functional form of the control function.

13.4.3 Method 3: IV-Probit

The last method we illustrate is IV-Probit. It has very strong model structure assumptions. Consider the following triangular system:

$$\begin{aligned} y_1^* &= z_1\delta_1 + \alpha_1 y_2 + u_1, \\ y_2 &= z_1\delta_{21} + z_2\delta_{22} + v_2, \\ y_1 &= \mathbf{1}(y_1^* > 0), \end{aligned} \tag{13.16}$$

where y_1^* is the latent utility, y_2 is the endogenous variable, z_1 is an exogenous control, z_2 is the instrument with $(u_1, v_2) \perp z = (z_1, z_2)$, and (u_1, v_2) is bivariate normal. The correlation between u_1 and v_2 is what makes y_2 endogenous.

The bivariate normality is the engine: it lets us write down the likelihood explicitly. Factor the joint density of the observed variables as

$$f(y_1, y_2 \mid z) = f(y_1 \mid y_2, z)\, f(y_2 \mid z), \tag{13.17}$$

where $f(y_2 \mid z)$ is the normal density of the first stage, and the conditional choice probability takes the Probit form

$$P(y_1 = 1 \mid y_2, z) = \Phi\left[\frac{z_1\delta_1 + \alpha_1 y_2 + (\rho_1/\tau_2)\,(y_2 - z\delta_2)}{\left(1 - \rho_1^2\right)^{1/2}}\right], \tag{13.18}$$

where $y_2 - z\delta_2$ is the first stage error (here $\delta_2 = (\delta_{21}', \delta_{22}')'$ stacks the first stage coefficients), τ_2 its standard deviation, and ρ_1 the correlation between the two errors. Conditioning on y_2 shifts the mean of u_1 by the part explained by v_2 (the control func-

tion logic again, in exact parametric form) and shrinks its variance by $1 - \rho_1^2$. We can then use MLE to jointly estimate all the parameters. One good thing is that Stata has the ready made package `ivprobit`, so implementation is one line.

13.5 Conclusion

This chapter completed the discrete choice toolkit. On estimation: log-linearization is quick but fragile (zero shares, absorbed parameters, inefficiency); MLE is the standard, and it is solved by numerical optimization, with a gradient-based updating rule and the Hessian as its speed control. On endogeneity, the main takeaways:

Main Takeaways: IV in the DCM

- Do not naively use the IV method from the linear model (2SLS) to solve an endogeneity issue in a non-linear model.
- You can use BLP, the control function, or IV-Probit.
- BLP fits the Logit model, but needs the endogeneity to happen at a higher level, the product-market level in the consumers' problem: pack it into fixed effects, estimate, then unpack with linear IV.
- The control function is pretty general, but you need to specify and flexibly estimate the control function itself.
- IV-Probit is simple to use with the Stata package, but it carries strong functional form assumptions: a linear triangular system with bivariate normal errors.

Beyond the specific methods, this chapter introduced two ideas with long lives in structural economics: the likelihood as the bridge from model to data, and the contraction mapping with its fixed point as the bridge from model to computation. Both will return whenever you estimate or solve a structural model in your future studies. This textbook does not aim to go deep into structural econometrics. For interested readers, please refer to related field courses.

Homework Problems

1. **Log-linearizing the Logit.**

Start from the Logit choice probability $P_{ni} = e^{V_{ni}} / \sum_j e^{V_{nj}}$.

(a) Take logs and derive the log-linearized equation, paying attention to the sign of the log sum term. Explain why the log sum can be absorbed by a fixed effect at the individual (or market) level when the data are grouped choices.

(b) List the situations in which this OLS approach breaks down or loses information, and explain each in one or two sentences: observed zero shares, parameters on individual level variables, and efficiency.

2. **One Newton–Raphson Step.**

Consider a log likelihood that is exactly quadratic in a scalar parameter,

$$LL(\theta) = a + b\,\theta - \frac{c}{2}\,\theta^2, \qquad c > 0.$$

(a) Compute the gradient g_t and the Hessian H_t at an arbitrary point θ_t.

(b) Perform one Newton–Raphson step with $\lambda = 1$, and show that θ_{t+1} equals the exact maximizer of LL, no matter where θ_t started. Explain in one sentence why NR solves a quadratic problem in a single step.

(c) Explain what can go wrong with the NR step when LL is not globally concave, and connect your answer to the sign of the Hessian.

3. **The Berry Update as a Fixed Point Iteration.**

Consider the Berry contraction update for the product-market fixed effects,

$$\delta_{jm}^{t+1} = \delta_{jm}^{t} + \ln S_{jm} - \ln \hat{S}_{jm}(\delta^t),$$

where S_{jm} are observed shares and $\hat{S}_{jm}(\delta)$ are the model's predicted shares.

(a) Show that a vector δ^* satisfying $\hat{S}_{jm}(\delta^*) = S_{jm}$ for all (j, m) is a fixed point of the update.

(b) Explain the economic logic of the update: what happens to δ_{jm} in the next iteration when the model underpredicts product j's share, and why is that the right direction?

(c) Using the Banach fixed point theorem, explain why, once the update is

known to be a contraction mapping, (i) the fixed point is unique and (ii) the iteration converges to it from any initial guess. Why is property (ii) important in practice?

Practice Example: Berry Inversion, Linear IV for Demand, and IV-Probit

In this practice example we implement the BLP logic in its simplest form, plain Logit demand without random coefficients, where the inversion from shares to fixed effects has a closed form, and we verify the Berry contraction numerically. We then estimate a triangular binary choice model with `ivprobit`. The example demonstrates the chapter's three lessons: endogeneity biases naive demand estimation, packing and unpacking the fixed effects restores the linear IV logic, and the contraction converges to the inversion from any starting point.

Data generating process. Two designs.

1. *Aggregate Logit demand across markets.* $M = 100$ markets, each with $J = 3$ inside products plus an outside option. Product attributes and shocks:

$$x_{jm} \sim \text{Uniform}(0, 2), \qquad w_{jm} \sim \text{Uniform}(0, 1), \qquad \xi_{jm} \sim \mathcal{N}(0, 0.25),$$

where the variance 0.25 corresponds to a standard deviation of 0.5, and where w_{jm} is a cost shifter (the instrument) and ξ_{jm} the unobserved quality. Prices are endogenous:

$$p_{jm} = 1 + 0.3\, x_{jm} + 1.0\, w_{jm} + 0.9\, \xi_{jm}.$$

Mean utilities and market shares follow the plain Logit model,

$$\delta_{jm} = 1 + 1.0\, x_{jm} - 2.0\, p_{jm} + \xi_{jm}, \qquad S_{jm} = \frac{e^{\delta_{jm}}}{1 + \sum_{j'=1}^{J} e^{\delta_{j'm}}},$$

with the outside share $S_{0m} = 1/(1 + \sum_{j'} e^{\delta_{j'm}})$.

2. *Triangular model for IV-Probit.* $N = 5{,}000$ individuals with exogenous control $z_1 \sim \mathcal{N}(0, 1)$ and instrument $z_2 \sim \mathcal{N}(0, 1)$. The first stage is $y_2 = 0.5z_1 + 1.0z_2 + v_2$, the latent utility is $y_1^* = 0.5z_1 + 1.0y_2 + u_1$, and (u_1, v_2) is bivariate normal with

unit variances and correlation 0.6; $y_1 = \mathbf{1}(y_1^* > 0)$.

Analysis steps. The do-file (`src/chapter 13/ch13_blp.do`) performs:

1. Simulate the market data and compute the exact model shares;
2. Recover the mean utilities by the closed-form inversion of Berry (1994), $\hat{\delta}_{jm} = \ln S_{jm} - \ln S_{0m}$, and confirm it reproduces the true δ_{jm};
3. Run OLS of $\hat{\delta}_{jm}$ on x_{jm} and p_{jm}: the price coefficient is biased toward zero, because high ξ products have both high prices and high demand;
4. Run the linear IV regression instrumenting p_{jm} with the cost shifter w_{jm}: the price coefficient recovers the truth -2, which is the BLP Step 2;
5. Implement the Berry contraction $\delta^{t+1} = \delta^t + \ln S - \ln \hat{S}(\delta^t)$ starting from $\delta^0 = 0$ and again from $\delta^0 = 5$, stopping when the sup norm of the update falls below 10^{-14} (at most 500 iterations), and verify both runs land on the closed-form inversion (maximum absolute deviation near machine precision);
6. Simulate the triangular design; show that a naive Probit of y_1 on (z_1, y_2) overstates α_1, while `ivprobit` recovers the truth; key lessons summary.

Key lessons.

- With Logit demand, market shares invert to mean utilities in closed form, and demand estimation becomes a linear regression of inverted shares on attributes and price.
- The endogeneity of price lives in that linear equation, so the standard IV toolkit applies: this is the pack, estimate, unpack logic of BLP in its simplest case.
- The Berry contraction finds the same inversion numerically, converging from any initial guess, exactly as the Banach fixed point theorem promises.
- When the model is a triangular system with normal errors, `ivprobit` handles the endogenous regressor in one line, at the price of strong functional form assumptions.

The do-file is located at `src/chapter 13/ch13_blp.do` and the simulated datasets at `data/chapter 13/`: `ch13_blp_markets.dta` (the aggregate Logit demand design) and `ch13_ivprobit.dta` (the triangular design).

Chapter 14

Peer Effect and Spillover

14.1 Introduction

In this final chapter, we investigate an important empirical question: how do we identify and estimate peer effects, or spillover effects? People may think it is straightforward and simple. Your outcome depends on your peers, so just run y on the peer average $\bar{y}$ or on the peer characteristics $\bar{x}$, and get the coefficient. But actually it is very complicated, and very dangerous.

We will discuss the issue from two perspectives:

- **Technically:** identification and inference failures. Some peer regressions are unidentified, and some produce coefficients that are mechanical artifacts of the regression itself.
- **Intuitively:** the interpretation of the peer effect coefficient. Even when a regression runs and delivers a significant number, what that number measures is usually not what the researcher wanted.

The related chapter of Angrist and Pischke (2009) is Section 4.6.2, but it is not so detailed; I recommend reading the original papers, Angrist (2014) and Manski (1993).

Let me give you a brief preview of the conclusions of this chapter:

1. You can never distinguish among endogenous effects, exogenous effects, and correlated effects. This is the *reflection problem*.
2. Never run regressions like y on $\bar{y}$ for the same group.

3. When running y on $\bar{x}$: make sure the group formation is random or quasi-random, and check all the alternative channels that can drive the result, such as measurement error.

4. Separate the people affecting others from the people being affected: the group generating $\bar{y}$ and $\bar{x}$ should be different from the group whose y you study.

The rest of the chapter unpacks these four statements one by one.

14.2 The Reflection Problem

14.2.1 Three Reasons Group Members Behave Alike

Peer effects are intrinsically very difficult to identify, because it is hard to distinguish among behavior causation, characteristics causation, and common environment. Manski (1993) named this the *reflection problem*, with a memorable image: when you see the co-movements of a person and his image in a mirror, without knowledge of optics, how can you tell whether the person's movements cause the movements of the image, or some external stimulus causes the person and the image to move together?

In general, individuals in the same group tend to behave similarly for three reasons:

- **Endogenous effects:** an individual's behavior is affected by the *behaviors* of the group;
- **Exogenous (contextual) effects:** an individual's behavior is affected by the exogenous *characteristics* of the group;
- **Correlated effects:** individuals in the same group have similar characteristics or face the same institutional environment.

Let us take students in a classroom as an example. Why do we see similar bullying behavior among students in the same class?

- *Endogenous effect:* a student bullies others because his or her friends do so;
- *Exogenous (contextual) effect:* a student bullies others because his or her friends come from violent families;
- *Correlated effect, version 1:* all students in this class bully others because they all come from families with violent fathers;

- *Correlated effect, version 2:* students in this class bully others because their head teacher does not care.

The endogenous and exogenous effects are different types of genuine spillovers: something about the peers reaches the individual. The correlated effect is purely a contamination: no causation flows between students at all; they simply share causes. Unfortunately, it is generally impossible to identify these three effects separately, even in a random or quasi-random environment. Let us see why.

14.2.2 The Linear-in-Means Model

Denote by y a scalar outcome, for example a student's test score; x a group attribute, for example the class indicator; z observed attributes that directly affect y, for example family SES; and u unobserved attributes that directly affect y, for example teacher ability. Consider the following equation:

$$y = \alpha + \beta\, E(y \mid x) + E(z \mid x)'\gamma + z'\eta + u, \tag{14.1}$$

where we assume $E(u \mid x, z) = x'\delta$, a CIA style quasi-random setting in which the unobserved terms can be absorbed into group fixed effects. In this notation, β is the *endogenous* effect (your group's mean outcome affects you), γ is the *exogenous* effect (your group's mean characteristics affect you), η is the direct effect of your own characteristics, and δ is the *correlated* effect (the group environment moves everyone).

Take the conditional expectation of (14.1) with respect to x and z:

$$E(y \mid x, z) = \alpha + \beta\, E(y \mid x) + E(z \mid x)'\gamma + z'\eta + x'\delta. \tag{14.2}$$

Can we identify β, γ, and δ separately?

14.2.3 The Social Equilibrium and the Counting Failure

Observe that the conditional expectation of y appears on *both sides*: the group mean outcome is itself an equilibrium object determined within the model. Take the expectation of (14.2) with respect to z, conditional on x:

$$E(y \mid x) = \alpha + \beta\, E(y \mid x) + E(z \mid x)'\gamma + E(z \mid x)'\eta + x'\delta. \tag{14.3}$$

Solving this "social equilibrium" equation for $E(y \mid x)$:

$$E(y \mid x) = \frac{\alpha}{1-\beta} + E(z \mid x)' \frac{\gamma + \eta}{1-\beta} + x' \frac{\delta}{1-\beta}. \tag{14.4}$$

Inserting (14.4) back into (14.2):

$$E(y \mid x, z) = \frac{\alpha}{1-\beta} + E(z \mid x)' \frac{\gamma + \beta\eta}{1-\beta} + x' \frac{\delta}{1-\beta} + z'\eta. \tag{14.5}$$

Equation (14.5) is the reduced form that the data can reveal. Using a linear regression, we can identify four objects separately: the composite intercept $\alpha/(1-\beta)$, the composite coefficient $(\gamma + \beta\eta)/(1-\beta)$ on the group mean characteristics, the composite coefficient $\delta/(1-\beta)$ on the group indicators, and the own effect η. But that is it; nothing more can be done. We have four regression coefficients and *five* unknowns $(\alpha, \beta, \gamma, \eta, \delta)$. Can we distinguish between the three effects? **No.**

The Reflection Problem (Manski, 1993)

In the linear-in-means model, the data identify only composites of the structural parameters: four regression coefficients against five unknowns. In general, we cannot distinguish between the endogenous effect β, the exogenous effect γ, and the correlated effect δ. This holds even under random or quasi-random group assignment: randomization fixes the correlated effects contamination coming from sorting, but the counting failure between β and γ remains, because the group mean outcome is an equilibrium object generated by the same model that we are trying to estimate.

This is disappointing. Can we still identify some meaningful spillover effect? The only hope is to give up on decomposing everything, and instead identify some simple *composite* effect: ignore the effect of $\bar{y}$ when running y on $\bar{x}$, or consider only y on $\bar{y}$ but not $\bar{x}$. However, even in this reduced ambition, we have to be very careful. Let us go to Angrist (2014) to see why.

14.3 The Perils of Peer Effects: Regressing y on $\bar{x}$

We have shown that carefully distinguishing the different peer effects is not feasible. Can we identify either the endogenous or the exogenous peer effect, treating the other

as a "channel"? That is, we run y only on $\bar{y}$ or only on $\bar{x}$, rather than both. What is the interpretation of these coefficients? It is not as straightforward as you may think. There are two kinds of peer effect regressions, and we discuss them one by one: this section takes the exogenous version, y on $\bar{x}$, and the next section takes the endogenous version, y on $\bar{y}$.

14.3.1 The Social Returns Regression and Its Anatomy

A good example of the exogenous effect is the *social return to education*: what is the impact of the province level average education on an individual's wage? We can directly run the regression

$$Y_{ij} = \mu + \gamma_0 s_i + \gamma_1 \bar{S}_j + \epsilon_{ij}, \tag{14.6}$$

where Y_{ij} is the wage of individual i in province j, s_i is the education of individual i, and $\bar{S}_j$ is the average education in province j. The coefficient of interest is γ_1, the spillover.

Angrist (2014) shows that the two coefficients of (14.6) can be expressed exactly as

$$\gamma_0 = \phi_1 + \psi(\phi_0 - \phi_1), \tag{14.7}$$

$$\gamma_1 = \psi(\phi_1 - \phi_0), \tag{14.8}$$

where the three ingredients are:

- ϕ_0 is the regression coefficient from a simple OLS regression of Y_{ij} on s_i;
- ϕ_1 is the coefficient from a 2SLS regression of Y_{ij} on s_i, using the group dummies $I(j)$ as instruments;
- $\psi = 1/(1 - R^2) > 1$ is a scalar related to the first stage R^2 of that 2SLS, the share of the variance of s_i explained by the group dummies.

Deriving (14.7) and (14.8) is Homework Problem 2.

How should we interpret this? We care about the spillover γ_1, and (14.8) says it is proportional to ψ and to the gap $(\phi_1 - \phi_0)$. So as long as $(\phi_1 - \phi_0) \neq 0$, we will have a nonzero γ_1. In words: as long as there is *any* difference between the estimate of an OLS regression of Y_{ij} on s_i and the estimate of a 2SLS regression of Y_{ij} on s_i using group dummies as instruments, you will find a nonzero "peer effect" in regression (14.6). The peer coefficient is not a new fact about the world; it is a repackaging of the OLS versus

IV gap.

14.3.2 When the Repackaging Is Meaningful, and When It Is Not

How can we understand this in our context? Suppose different provinces are randomly assigned compulsory education laws. The OLS regression pins down the correlation between your own wage and your own education *within* the sample, comparing you with people in your own province among others. If peer effects exist, this OLS understates the total (group level) effect of education $\phi_1 = \gamma_0 + \gamma_1$: when individual i's education rises, the spillover also raises the wages of the untreated people in the same province, who serve as i's comparison group, so the within province comparison captures only the private part and misses the part shared with the peers. Meanwhile, the IV regression essentially compares outcomes *across* provinces whose education variation is driven by the randomly assigned laws; this comparison is not contaminated by the spillover, because the spillover happens within provinces. Subtracting OLS from IV, and scaling by ψ, then recovers the spillover: this is the sense in which (14.8) is a real identification result under ideal conditions.

But is spillover the only reason why the IV result can deviate from OLS? **Of course not.** There can be many reasons for a difference between OLS and 2SLS estimates: selection bias, measurement error, and so on. If selection bias exists, OLS can overestimate the private return, generating a gap with the opposite sign. If education is measured with classical error, OLS suffers attenuation bias while the grouped 2SLS largely does not, because averaging within groups washes out idiosyncratic measurement error; this alone creates a gap and hence a spurious "peer effect."

A Significant Peer Coefficient Is Not Evidence of Peer Effects

Peer effects are not essential for the existence of the OLS versus IV difference. Even if you detect a nonzero and significant γ_1 in the social returns regression (14.6), it can be entirely due to selection bias or measurement error. And even if a real peer effect exists, the estimate of γ_1 is contaminated by these same forces. The regression cannot tell you which story you are in.

14.3.3 A Diagnostic for Measurement Error

Do we have any method to alleviate this issue? For selection bias, not much can be done beyond design. But we *can* test whether a detected "peer effect" actually comes from measurement error, using a simulation procedure from Carrell, Hoekstra, and Kuka (2018) and Feld and Zölitz (2017).

The basic idea is simple. We worry that the detected γ_1 in (14.6) is produced by measurement error in the education variable. A direct implication is that if we *create more* measurement error, the magnitude of γ_1 should grow. Assume we proxy education by college degree attainment, and that the sample of size N has x percent college educated. The procedure is:

1. Randomly select p percent of the sample;
2. Within the selected subsample, randomly assign x percent of individuals to have college education, replacing their true education status in the data;
3. Run the main regression (14.6) on this partially faked data.

Repeat the process varying p from 0 to 100 percent: $p = 0$ is the baseline estimate without any added error, and $p = 100$ is the extreme case where the education variable is pure noise. If the estimated γ_1 grows larger and larger as more noise is added, then measurement error is likely an important driver of the detected coefficient. If it does not, then it is more plausible that γ_1 reflects a true peer effect rather than measurement error.

One important tip to close this section: **full randomization of group membership, as in an RCT, cannot solve this problem.** The issue is not about whether $\bar{S}_j$ is randomly assigned; it is about why the two auxiliary regressions, the OLS of Y_{ij} on s_i and the 2SLS of Y_{ij} on s_i with group dummies as instruments, can give different answers. Randomizing group formation does nothing to equalize them when selection into education or measurement error is present.

14.4 The Perils of Peer Effects: Regressing y on $\bar{y}$

We have discussed regressing y on $\bar{x}$. The second case is regressing y on $\bar{y}$, which seems to inform us about the endogenous effect. However, this regression is even more dangerous than the first one.

14.4.1 The Mechanical Coefficient of One

To begin with, directly running y on the full group mean $\bar{y}$ makes no sense: it will give you a coefficient of one, mechanically. Why? Consider a school dropout problem. Let s_{ij} be the dropout decision of student i in school j, and $\bar{S}_j$ the average dropout rate in school j. We run

$$s_{ij} = \mu + \phi_2 \bar{S}_j + \epsilon_{ij}. \tag{14.9}$$

The OLS will give you $\hat{\phi}_2 = 1$ for sure. Here is a simple proof. With school sizes n_j and overall mean $\bar{S}$,

$$\hat{\phi}_2 = \frac{\sum_j \sum_i s_{ij}(\bar{S}_j - \bar{S})}{\sum_j \sum_i (\bar{S}_j - \bar{S})^2} = \frac{\sum_j (\bar{S}_j - \bar{S}) \sum_i s_{ij}}{\sum_j n_j (\bar{S}_j - \bar{S})^2} = \frac{\sum_j (\bar{S}_j - \bar{S}) n_j \bar{S}_j}{\sum_j n_j (\bar{S}_j - \bar{S})^2},$$

and expanding the denominator,

$$\hat{\phi}_2 = \frac{\sum_j n_j \bar{S}_j (\bar{S}_j - \bar{S})}{\sum_j \left[n_j \bar{S}_j (\bar{S}_j - \bar{S}) - n_j \bar{S} (\bar{S}_j - \bar{S}) \right]} = 1,$$

where the second term of the denominator vanishes because $\sum_j n_j (\bar{S}_j - \bar{S}) = 0$ by the definition of the overall mean. The numerator and the denominator are identical. Nothing about behavior was used anywhere in this argument: the coefficient of one is pure accounting, because each s_{ij} is a component of its own regressor.

14.4.2 Leave-One-Out, and Why It Is Still Not Enough

To avoid this mechanical problem, we can run a *leave-one-out* regression:

$$s_{ij} = \mu + \phi_3 \bar{S}_{-i,j} + \tilde{\epsilon}_{ij}, \tag{14.10}$$

where $\bar{S}_{-i,j}$ is the average dropout rate in school j excluding student i himself. The coefficient of this regression is no longer guaranteed to be one. However, it is still almost impossible to claim that it identifies a peer effect or spillover, because *any school level random shock creates a spurious peer effect*. For example, a good principal leads all students in a school to stay enrolled: everyone's dropout falls together, your outcome and your peers' outcomes co-move, and the leave-one-out regression reports a strong positive "peer effect" that has nothing to do with spillovers between students. We are back to Manski (1993): it is almost impossible to distinguish real endogenous effects

from the contamination of correlated effects.

14.4.3 Randomization, Selection, and Variation

Beyond the issues above, we also need to be careful about the traditional selection problem. Usually, grouping is not random: good students select into good schools, good employees select into good firms. If group membership is chosen, then peer averages are correlated with everything that drove the choice. Thus, the prerequisite of any peer effect analysis, before the previous issues even arise, is a random or quasi-random group formation.

However, once you have random group formation, the *variation* of the independent variable becomes a problem. If students are randomly assigned to schools, then $E[s_{ij}]$ is the same for all schools; and if the number of students in each school is very large, the realized averages $\bar{S}_j$ will also be very similar across schools, by the law of large numbers. But you need variation in $\bar{S}_j$ to identify the peer effect. So there is a tradeoff. If the grouping is totally random, you may have very little variation in the independent variable $\bar{S}$. If the grouping is not that random, you may have plenty of variation in $\bar{S}$, but the selection issue can be severe. Therefore, in practice, the best case is: a genuinely random grouping, with group sizes that are *not too large*, so that small sample variation still gives you usable differences in $\bar{S}_j$ across groups. Usually, in randomization/quasi-randomization cases, we should show summary statistics to validate that we still have enough variation after randomization.

14.5 Empirical Suggestions and Application

In general, peer effects are difficult to identify. Here are some empirical suggestions.

Suggestion 1: clearly separate the subjects who receive the peer effect from the peers who provide it. Do not ask "what is the impact of fellow boys' test scores on a boy," where the outcome group and the peer group overlap and all the mechanical and reflection problems return. Ask instead "what is the impact of boys' test scores on a girl": the group generating the peer measure is disjoint from the group whose outcome is studied, which kills the mechanical feedback by construction.

Suggestion 2: make sure the underlying OLS and 2SLS would agree in the absence of peer effects. We cannot do much to guarantee this, but one thing you

should always do is check the measurement error channel using the noise injection simulation of Carrell, Hoekstra, and Kuka (2018) and Feld and Zölitz (2017) described in Section 14.3.

Suggestion 3: manage the tradeoff between randomization and variation. I would always put randomization as the most important thing. A balance check is therefore essential as the first step of any peer effect analysis: regress $\bar{S}$ on potential confounders to verify that the grouping is as good as random. Then check that you still have enough variation in $\bar{S}$ after randomization to identify anything.

A Checklist for Peer Effect Papers

1. Is the peer group disjoint from the outcome group? If not, expect mechanical bias and the reflection problem.
2. Is group formation random or quasi-random? Run the balance check: $\bar{S}$ on predetermined confounders.
3. Is there residual variation in $\bar{S}$ after randomization? Small or moderate group sizes help.
4. Could measurement error alone produce the coefficient? Run the noise injection diagnostic.
5. Could a group level shock alone produce the coefficient? Remember that correlated effects survive every regression fix.

14.6 Application: School Restrictions, Migration, and Peer Effects

The application paper for this week is Huang and Zhang (2023), which investigates the impact of migrant children's school enrollment restrictions on education outcomes in China. The paper has two parts: a *peer effect estimation* of migrant and left-behind children on their classmates, which confronts exactly the identification issues of this chapter, and a *spatial equilibrium model* used to evaluate the enrollment policy, built on the discrete choice tools of Chapters 12 and 13. It is therefore the natural capstone for the last three chapters of this textbook, and we describe it in some detail.

Research question and background. Under China's Hukou household registration system, children without a local Hukou can be denied enrollment in public schools

at their parents' migration destination. Migrant parents therefore face a hard choice: take the children along and risk relegation to low quality private migrant schools, or leave the children behind in the hometown where public school access is guaranteed but parents are absent. The result is two disadvantaged groups, *migrant children* and *left-behind children*, which together numbered 103 million in 2015. Local parents in developed cities resist relaxing the restriction, fearing negative spillovers onto their own children; local governments fear the fiscal cost. The paper asks whether these two fears are justified: do children of migrant families really generate negative spillovers on classmates, and, accounting for peer effects in general equilibrium, would relaxing the restriction raise national human capital at an acceptable cost?

Peer effect estimation. The first part estimates the spillovers of migrant and left-behind classmates using the China Education Panel Survey, and the identification design is a direct application of this chapter's playbook. The regression is of the exogenous composite type: student cognitive test score on the *leave-one-out* proportions of migrant and left-behind peers in the class, with school fixed effects. Random group formation comes from an institutional feature: Chinese law forbids middle schools from assigning students to classes by ability or family background, and the paper keeps only schools where both the principal and the teachers report random assignment. Randomization is then *verified*, not assumed, by the balance check this chapter recommends: within schools, class level peer proportions are uncorrelated with student, family, and teacher characteristics. The variation side of the tradeoff from Section 14.4.3 is handled by moderate group sizes, about seven classes of fifty students per school. The results: a ten percentage point increase in the proportion of left-behind peers lowers classmates' test scores by about 0.088 standard deviations, while the same increase in migrant peers costs a modest 0.038 standard deviations; in the second year the migrant effect vanishes and the left-behind effect shrinks but persists. The robustness appendix implements both checks this chapter emphasized: the regressions are rerun with only ordinary local students as subjects (separating receivers from providers), and the noise injection simulation of Carrell, Hoekstra, and Kuka (2018) and Feld and Zölitz (2017) shows the estimates *attenuating toward zero* as noise is added, the signature of a real effect rather than a measurement error artifact.

The spatial equilibrium model and policy results. A regression alone cannot answer the policy question, because relaxing the restriction changes who migrates, where children study, and hence every classroom's peer composition. The second part

therefore embeds the estimated peer effects in an Eaton and Kortum style spatial equilibrium model (Eaton and Kortum, 2002), built from the discrete choice tools of Chapters 12 and 13: parents' city choice follows a Fréchet taste draw that delivers a gravity equation, the child migration decision is a binary Logit whose expected value enters parents' utility through the log-sum formula, and the several hundred city fixed effects in children's human capital are recovered with the contraction mapping algorithm of Berry, Levinsohn, and Pakes (1995). In the main counterfactual, removing the enrollment restriction increases worker migration, draws about 1.26 million new migrant students into cities, and reunites families. The directly treated children who move from migrant schools into public schools gain a remarkable 0.76 standard deviations of human capital, and national average human capital rises even after the short run negative spillovers on big city children are charged against it. The fiscal cost is about 30 billion RMB, only 1.5 percent of annual government education expenditure, and the gains to parents and children are about twice the cost.

This paper helps you understand how to apply what we learned in the last three chapters together. The peer effect regressions survive scrutiny precisely because they follow the checklist of Section 14.5, and the policy question is answerable only because the discrete choice machinery of Chapters 12 and 13 turns those regression estimates into a general equilibrium answer.

14.7 Conclusion

Main Takeaways on Peer Effects

- The reflection problem is fundamental: endogenous, exogenous, and correlated effects cannot be separately identified, even under randomization.
- Never regress y on the full group mean $\bar{y}$: the coefficient is one by accounting, not by behavior. Leave-one-out removes the mechanical bias but not the correlated effects.
- A significant coefficient on $\bar{x}$ is a repackaged OLS versus IV gap: $\gamma_1 = \psi(\phi_1 - \phi_0)$. Selection and measurement error produce the same gap that true spillovers do; run the noise injection diagnostic.
- Randomize group formation first, check balance, and keep groups small enough that $\bar{S}$ still varies.

- Separate the receivers of the effect from the providers of the effect.

This chapter closes the textbook. We began with potential outcomes and a single binary treatment in the first chapter; we end with social interactions, where each agent's outcome feeds into every other agent's regressor, and with structural models, where behavior itself is the object of estimation. The common discipline throughout has been the same: know your estimand, know the thought experiment behind your standard error, and know what your regression is actually averaging before you attach a story to it.

Homework Problems

1. **The Reflection Problem as a Counting Failure.**

   Consider the linear-in-means model (14.1) with $E(u \mid x, z) = x'\delta$:

   $$y = \alpha + \beta\, E(y \mid x) + E(z \mid x)'\gamma + z'\eta + u.$$

   (a) Derive the social equilibrium equation (14.4) by taking expectations with respect to z and solving for $E(y \mid x)$.

   (b) Substitute back to obtain the reduced form (14.5), and list the four coefficient composites that a linear regression can identify.

   (c) Explain, in terms of counting equations and unknowns, why β, γ, and δ cannot be separately identified, and why random assignment of individuals to groups does not repair this.

2. **The Anatomy of the Social Returns Regression.**

   Consider the regression $Y_{ij} = \mu + \gamma_0 s_i + \gamma_1 \bar{S}_j + \epsilon_{ij}$, where $\bar{S}_j$ is the group mean of s_i. Let ϕ_0 be the coefficient from the OLS of Y_{ij} on s_i, let ϕ_1 be the coefficient from the 2SLS of Y_{ij} on s_i using group dummies as instruments, and let R^2 be the first stage R squared of that 2SLS, with $\psi = 1/(1 - R^2)$.

   (a) Decompose $s_i = \bar{S}_j + (s_i - \bar{S}_j)$ and argue that the two components are orthogonal. Show that the regression of Y on s_i and $\bar{S}_j$ is equivalent to a regression of Y on the within component $(s_i - \bar{S}_j)$ and the between component

$\bar{S}_j$, and relate the coefficients of the two regressions.

(b) Show that the between coefficient equals ϕ_1 and that the OLS coefficient ϕ_0 is a variance weighted average of the within and between coefficients, with weight $(1 - R^2)$ on the within component.

(c) Combine (a) and (b) to derive $\gamma_0 = \phi_1 + \psi(\phi_0 - \phi_1)$ and $\gamma_1 = \psi(\phi_1 - \phi_0)$, and verify that $\gamma_0 + \gamma_1 = \phi_1$.

(d) Give two reasons other than peer effects why $\phi_1 \neq \phi_0$, and state in one sentence what this implies for interpreting γ_1.

Practice Example: Mechanical, Spurious, and Fake Peer Effects

In this practice example we manufacture every pathology of this chapter in a controlled laboratory: the mechanical coefficient of one, the spurious peer effect from a group level shock, and the fake social return produced by measurement error, together with the noise injection diagnostic that detects it. In all designs the *true* peer effect is known, so every estimated "peer effect" can be compared against the truth.

Data generating process. Two designs.

1. *Classrooms with random assignment (mechanical and spurious effects).* $J = 200$ classes with $n = 20$ students each (4,000 students), randomly assigned. The dropout indicator is $s_{ij} = \mathbf{1}(u_{ij} < 0.2)$ with $u_{ij} \sim \text{Uniform}(0, 1)$, i.i.d. across all students: *no group structure and no peer effect.* A second outcome carries a class level shock but again no peer effect:

$$y_{ij} = 1 + v_j + e_{ij}, \qquad v_j \sim \mathcal{N}(0, 0.25), \quad e_{ij} \sim \mathcal{N}(0, 1).$$

   A predetermined family background covariate $f_{ij} \sim \mathcal{N}(0, 1)$ is also drawn and saved, so that a balance check can be run as an exercise.

2. *Provinces (fake versus true social returns).* $J = 100$ provinces with $n = 50$ workers each (5,000 workers). In both scenarios the observed education measure is the college proxy $c_{ij} = \mathbf{1}(s_{ij} \geq 14)$ applied to the scenario's education variable, the own return is 0.10, and the wage noise is $u_{ij} \sim \mathcal{N}(0, 0.25)$. In the *fake* scenario,

true education clusters within provinces and wages carry *no* social return:

$$s_{ij}^{fake} = 12 + a_j + e_{ij}, \quad a_j \sim \mathcal{N}(0,1), \quad e_{ij} \sim \mathcal{N}(0,4), \quad Y_{ij} = 1 + 0.10\, s_{ij}^{fake} + u_{ij};$$

any estimated peer coefficient on the province college share $\bar{C}_j$ is produced by the interaction of measurement error (the coarse college proxy) with the clustering of true education. In the *true* scenario, education is randomly distributed across provinces with the same total variance and wages carry a genuine social return of 0.50 through the college share:

$$s_{ij}^{true} = 12 + e_{ij}, \qquad e_{ij} \sim \mathcal{N}(0,5), \qquad Y_{ij} = 1 + 0.10\, s_{ij}^{true} + 0.50\, \bar{C}_j + u_{ij};$$

here the identifying variation in $\bar{C}_j$ is exactly the small sample variation under random assignment discussed in Section 14.4.3.

Analysis steps. The do-file (`src/chapter 14/ch14_peer.do`) performs:

1. Simulate Design 1; regress s_{ij} on the full class mean $\bar{S}_j$ and confirm the coefficient is exactly one (the mechanical result of Section 14.4);
2. Regress s_{ij} on the leave-one-out mean $\bar{S}_{-i,j}$ and confirm the coefficient is near zero: leave-one-out removes the mechanical bias when there is truly no group structure;
3. Regress y_{ij} on the leave-one-out mean $\bar{y}_{-i,j}$ and confirm a large spurious "peer effect" near the population value $\sigma_v^2/(\sigma_v^2 + \sigma_e^2/(n-1)) \approx 0.83$ (about 0.79 in this simulated sample), manufactured entirely by the class level shock;
4. Simulate Design 2; run the social returns regression Y_{ij} on c_{ij} and $\bar{C}_j$ in both scenarios: the fake scenario shows a positive and significant "social return" despite the truth being zero, while the true scenario recovers the genuine coefficient 0.50 within sampling error (point estimate about 0.61, SE 0.14). For the fake scenario, decompose the artifact via $\gamma_1 = \psi(\phi_1 - \phi_0)$ by also reporting the OLS coefficient ϕ_0, the group dummy 2SLS coefficient ϕ_1, and the first stage R^2;
5. Run the noise injection diagnostic of Carrell, Hoekstra, and Kuka (2018) and Feld and Zölitz (2017) on both scenarios, varying p from 0 to 100 percent in steps of 10: record $\hat{\gamma}_1(p)$ (averaged over replications), plot the two paths, and read the diagnostic. The fake coefficient *grows* as noise is added, collapsing only when

the regressor approaches pure noise, while the true coefficient declines steadily toward zero as its peer measure is randomized away.

Key lessons.

- The full mean regression returns one by accounting; the leave-one-out regression repairs only the accounting, not the identification.
- A modest common shock manufactures a large, highly significant, and completely spurious peer effect in y on $\bar{y}_{-i}$ regressions.
- A fake social return arises from measurement error interacting with clustered characteristics, exactly the $\psi(\phi_1 - \phi_0)$ anatomy; the noise injection paths separate the fake from the true.
- Every pathology occurs with the truth known: significance is not identification.

The do-file is located at `src/chapter 14/ch14_peer.do` and the simulated datasets at `data/chapter 14/`.

Appendix A

Replication Guide

Every chapter ends with a practice example built on a simulated dataset. The simulation code, the saved datasets, and the figure generation scripts live in the companion repository, https://github.com/hzbbb/Notes-on-Applied-Econometrics, in three parallel folder trees: `src/chapter N/` holds the Stata do file (and, where a chapter has data based figures, a `notes figures/` subfolder with the Julia script that draws them), `data/chapter N/` holds the datasets the do file saves, and `Notes/Chapter N/` holds the chapter source.

Software. The do files were written for Stata 16 and later and verified in Stata 17; each begins with a `version` statement (Chapter 11 is pinned to version 17 because its shipped dataset reproduces exactly under that version). Chapters 3 and 9 use community contributed or recent built in commands (`lasso`, `elasticnet`, `twowayfeweights`); the do files say where to obtain them. Figures generated from simulated data use Julia 1.12 with the `Plots`, `Distributions`, and `StatsBase` packages.

Reproducibility. Every do file sets its random number seed once, before the first random draw, and draws all random variables before any sort that depends on them, so the saved datasets are a deterministic function of the seed. Running a do file from its own folder regenerates the datasets and, where applicable, the exported figures. The table lists, for each chapter, the do file, the seed(s), the datasets it saves, and the figure script.

Chapter	Do file	Seed(s)	Datasets	Figure script
1	`ch1_rct.do`	20250912	`ch1_rct.dta`	
2	`ch2_nonpar.do`	20250919	`ch2_nonpar.dta`	generate_figures.jl
3	`ch3_ml.do`	20250926	`ch3_ml.dta`	generate_figures.jl
		20250927		
4	`ch4_dag.do`	20251017	`ch4_dag.dta`	
5	`ch5_iv.do`	20251024	`ch5_iv.dta`	
6	`ch6_iv_beyond_late.do`	20251026	`ch6_iv_beyond_late.dta`	generate_figures.jl
7	`ch7_bartik.do`	20251107	`ch7_bartik.dta`	
8	`ch8_did.do`	20251122	`ch8_did.dta`	generate_figures.jl
9	`ch9_staggered_did.do`	20251121	`ch9_staggered_A.dta`	
			`ch9_staggered_B.dta`	
10	`ch10_rd.do`	20251202	`ch10_rd_fuzzy.dta`	generate_figures.jl
			`ch10_rd_sharp.dta`	
11	`ch11_cluster.do`	20251203	`ch11_cluster.dta`	
12	`ch12_logit.do`	20251214	`ch12_logit_binary.dta`	
			`ch12_logit_modes.dta`	
13	`ch13_blp.do`	20251226	`ch13_blp_markets.dta`	
			`ch13_ivprobit.dta`	
14	`ch14_peer.do`	20260102	`ch14_classrooms.dta`	
			`ch14_provinces.dta`	

Homework solution keys. The solution keys to the end of chapter problems are kept out of this book; they are maintained as separate documents for instructors in the same repository.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association*, 105(490): 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science*, 59(2): 495–510.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica*, 88(1): 265–296.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *Quarterly Journal of Economics*, 138(1): 1–35.

Abadie, Alberto. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature*, 59(2): 391–425.

Adao, Rodrigo, Michal Kolesár, and Eduardo Morales. 2019. "Shift-share Designs: Theory and Inference." *Quarterly Journal of Economics*, 134(4): 1949–2010.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*, 80(3): 313–336.

Angrist, Joshua D. 2014. "The Perils of Peer Effects." *Labour Economics*, 30: 98–108.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review*, 103(6): 2121–2168.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica*, 63(4): 841–890.

Berry, Steven T. 1994. "Estimating Discrete-Choice Models of Product Differentiation." *RAND Journal of Economics*, 25(2): 242–262.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119(1): 249–275.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. "Quasi-experimental Shift-share Research Designs." *Review of Economic Studies*, 89(1): 181–213.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting event-study designs: robust and efficient estimation." *Review of Economic Studies*, 91(6): 3253–3285.

Callaway, Brantly and Pedro H.C. Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with a Continuous Treatment." arXiv preprint arXiv:2107.02637.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica*, 82(6): 2295–2326.

Card, David and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review*, 84(4): 772–793.

Card, David and Alan B. Krueger. 2000. "Minimum Wage and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply." *American Economic Review*, 90(5): 1397–1420.

Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2015. "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica*, 83(6): 2453–2483.

Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2017. "Regression Kink Design: Theory and Practice." In *Regression Discontinuity Designs: Theory and Applications* (Advances in Econometrics, Vol. 38), Emerald Publishing, 341–382.

Card, David. 2009. "Immigration and Inequality." *American Economic Review*, 99(2): 1–21.

Carrell, Scott E., Mark Hoekstra, and Elira Kuka. 2018. "The Long-Run Effects of Disruptive Peers." *American Economic Review*, 108(11): 3377–3415.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6B:5549–5632.

Clark, Andrew E., Huifu Nong, Hongjia Zhu, and Rong Zhu. 2021. "Compensating for Academic Loss: Online Learning and Student Performance during the COVID-19 Pandemic." *China Economic Review*, 68: 101629.

de Chaisemartin, Clément and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, 110(9): 2964–2996.

Dingel, Jonathan I. and Felix Tintelnot. 2020. "Spatial Economics for Granular Settings." NBER Working Paper 27287.

Eaton, Jonathan and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica*, 70(5): 1741–1779.

Feld, Jan and Ulf Zölitz. 2017. "Understanding Peer Effects: On the Nature, Estimation, and Channels of Peer Effects." *Journal of Labor Economics*, 35(2): 387–428.

Gelman, Andrew and Guido Imbens. 2019. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics*, 37(3): 447–456.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik Instruments: What, When, Why, and How." *American Economic Review*, 110(8): 2586–2624.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*, 69(1): 201–209.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.

He, Guojun, Shaoda Wang, and Bing Zhang. 2020. "Watering Down Environmental Regulation in China." *Quarterly Journal of Economics*, 135(4): 2135–2185.

Heckman, James J. and Edward J. Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Heckman, J. J. and E. Vytlacil (2007). Econometric evaluation of social programs, part I: Causal models, structural models, and econometric policy evaluation. In J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6B, pp. 4779–4874. Elsevier.

Heckman, James J. and Edward J. Vytlacil. 2007a. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." *Handbook of Econometrics*, 6: 4779–4874.

Heckman, James J. and Edward J. Vytlacil. 2007b. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments." *Handbook of Econometrics*, 6: 4875–5143.

Huang, Zibin and Junsen Zhang. 2023. "School Restrictions, Migration, and Peer Effects: A Spatial Equilibrium Analysis of Children's Human Capital in China." Unpublished Manuscript.

Imbens, Guido W. and Joshua D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475.

Imbens, Guido and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies*, 79(3): 933–959.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Imbens, Guido W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*, 58(4):1129–1179.

Kline, Patrick and Christopher R. Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *Quarterly Journal of Economics*, 131(4): 1795–1848.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics*, 114(2): 497–532.

Lee, David S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics*, 142(2): 675–697.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–870.

Liang, Kung-Yee and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, 73(1): 13–22.

Lucas, R. E. Jr. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1, 19–46.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*, 60(3): 531–542.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. Oxford University Press.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics*, 142(2): 698–714.

McFadden, Daniel. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by Paul Zarembka, 105–142. Academic Press.

McFadden, Daniel. 1978. "Modeling the Choice of Residential Location." In *Spatial Interaction Theory and Planning Models*, edited by Anders Karlqvist et al., 75–96. North-Holland.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics*, 32(3): 385–397.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142.

Neal, Brady (2020). Introduction to Causal Inference. Course Lecture Notes (draft). Available at https://www.bradyneal.com/causal-inference-course.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*, 37(2): 187–204.

Pearl, Judea and Dana Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. Allen Lane.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. 2nd edition. Cambridge University Press.

Pinto, Rodrigo. 2015. "Selection Bias in a Controlled Experiment: The Case of Moving to Opportunity." Unpublished Ph.D. Thesis, University of Chicago, Department of Economics.

Rambachan, Ashesh and Jonathan Roth. 2023. "A More Credible Approach to Parallel Trends." *Review of Economic Studies*, 90(5): 2555–2591.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *Journal of Econometrics*, 235(2): 2218–2244.

Roth, Jonathan. 2022. "Pretest with Caution: Event-study Estimates after Testing for Parallel Trends." *American Economic Review: Insights*, 4(3): 305–322.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66(5), 688–701.

Sun, Liyang and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge University Press.

Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica*, 70(1): 331–341.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372.