Assumptions in Causal Inference: Illuminating the Path to Credibility

Suggested Citation: Xi Chen (XX), "Assumptions in Causal Inference: Illuminating the Path to Credibility", : Vol. xx, No. xx, pp 1–XX. DOI: XXXXXXXXX.

Xi Chen
Erasmus University Rotterdam
chen@rsm.nl

now
the essence of knowledge
Boston — Delft

Contents

Assumptions in Causal Inference: Illuminating the Path to Credibility

Xi Chen[1]

[1] *Erasmus University Rotterdam; chen@rsm.nl*

ABSTRACT

TBA

Xi Chen (XX), "Assumptions in Causal Inference: Illuminating the Path to Credibility", : Vol. xx, No. xx, pp 1–XX. DOI: XXXXXXXXX.

1

Introduction

In the vibrant field of marketing, marketing researchers are always open to new methodologies. It is well exemplified by the wide acceptance and fast integration of causal inference methods in marketing research. In leading academic journals, the number of articles using causal inference methods has been constantly increasing in the last decade[1]. This is not surprising, given that marketing has a long tradition of analyzing data to understand the effect of marketing decisions. Historical tools such as BRANDAID (Little, 1975) and ADVISOR (Lilien, 1979) were designed to help marketers quantify the effectiveness of marketing mixes so that marketing budget could be optimally allocated. As a result, in recent years, we have seen the take-off of causal inference methods in marketing research and the embracement of credibility revolution by marketing researchers.

For marketing researchers, the fundamental objective of embracing causal inference is to increase the credibility of their analysis. Many marketing problems are plagued by endogeneity issues (Papies *et al.*, 2023; Rutz and Watson, 2019), such that statistical analysis alone is not sufficient to address these problems. A case in point is the estimation

[1] See Figure 1 in the review article by Papies *et al.* (2023)

of the price effect. A well-known problem of "simultaneity" (that is, prices are determined simultaneously by demand and supply) prevents us from obtaining a credible estimate of the effect of price by regressing sales on prices (see e.g., Chintagunta, 2001). In such cases, marketing researchers turn to causal inference [2].

Despite the growing reliance on causal inference to enhance credibility in research, its fundamental aspects remain underexplored. Researchers often turn to causal inference methods as a means of ensuring credibility, yet there has been limited systematic discussion of why causal inference is considered more credible than purely statistical approaches and what factors determine its credibility.

This monograph therefore seeks to address these fundamental questions by examining the core principles of causal inference and the determinants of its credibility. Beyond theoretical discussions, it also aims to provide practical insights into how researchers can assess and enhance the credibility of their causal inference efforts. Ultimately, this monograph offers a systematic framework that illuminates the path to credibility in causal inference, equipping empirical researchers with the tools necessary to credibly apply these methods.

1.1 Scope and Main Theses

Before delving into the formal content, I first define the scope of my discussion and outline the key theses. Although many factors contribute to the credibility of causal inference methods, this work focuses on the fundamental principles that distinguish causal inference as a distinct field. Statistical factors, such as measurement accuracy, functional forms, and distributional assumptions, are, without a doubt, important for causal inference. However, they are not the main focus of this monograph as they are extensively discussed in the statistical literature.

This work prioritizes fundamental principles because they serve as the foundation for credible causal analysis. They must be carefully examined and properly established. Without a solid foundation, even sophisticated methodologies and rich data cannot ensure valid causal

[2] For example, Sudhir (2001) proposes instrumental variables based on competitor characteristics to address the price endogeneity problem.

conclusions. Next, I will introduce two key theses and further explain the reasoning behind this focus.

The first key thesis is that the distinctiveness of causal inference lies in its emphasis on identification, a formal axiomatization process that logically ensures researchers can infer causal estimands from data. By this line of reasoning, causal inference is grounded in a deductive framework that establishes (sufficient) conditions under which causal estimands can be learned. This notion of identification is fundamental for both Rubin's potential outcome framework (Rubin, 1974) and Pearl's causal graph approach (Pearl, 2009). For example, Rubin's framework introduces the Stable Unit Treatment Value Assumption (SUTVA), which ensures that the average treatment effect (ATE) can be validly estimated in a randomized experiment. Similarly, Pearl's do-calculus formalizes a set of conditions that equate conditional probabilities with an intervention (the `do`-operation). Based on the conditions, causal effects can then be inferred. Therefore, the axiomatization process underlying these approaches is fundamental to causal inference as a unique and rigorous discipline for causal analysis.

A case in point from marketing is the inference of the price effect. Suppose a researcher collects data on sales and prices and runs a regression $\text{Sales} = \beta\text{Price} + \varepsilon$ to learn β. If we take the conditional expectation of Price on the equation, we have

$$\text{E}(\text{Sales}|\text{Price}) = \beta\text{Price} + \text{E}(\varepsilon|\text{Price}) \tag{1.1}$$

The conditional expectation $\text{E}(\text{Sales}|\text{Price})$ can be directly learned from data. To know β, we must know $\text{E}(\varepsilon|\text{Price})$, which cannot be learned from data. Therefore, we make the so-called "zero-conditional mean assumption" to let $\text{E}(\varepsilon|\text{Price}) = 0$. Under this assumption, we can infer the causal estimand ("price effect") from the data. If this assumption does not hold, we cannot learn anything about the price effect even if the data have millions of observations.

The second key thesis is what I term "The Golden Formula of Empirical Analysis" (see Proposition 1.1). In empirical research, conclusions are derived from data to address specific research questions. However, data alone are often insufficient to generate meaningful conclusions. A key insight, as illustrated in the earlier example of the price effect, is

that empirical analysis must integrate both assumptions and data to produce conclusions (see Manski, 2013, p.11). Without well-founded assumptions, even the most comprehensive datasets may fail to provide conclusive answers to research questions.

The Golden Formula of Empirical Analyses

Assumptions + Data ⇒ Conclusions

An implication from Proposition 1.1 is that different sets of assumptions can result in different interpretations and managerial implications of the same causal estimand, even with the same data. A good example is the coupon field experiment (e.g., Bawa and Shoemaker, 1987), where the non-compliance issue arises because consumers who receive coupons may not redeem them. The sole assumption that the assignment of coupon is random (and SUTVA) identifies the intention-to-treat effect, which informs managers the effect of the campaign on sales. In comparison, the additional set of assumptions in the seminal paper by Imbens and Angrist (1994) identify the local average treatment effect (LATE), which informs managers the effect of coupon usage on sales for "compliers."

Consistent with Proposition 1.1, recent studies in various fields also show that scientists reach vastly different conclusions for the same research questions when given the same data, because they make different assumptions in their analysis (Botvinik-Nezer *et al.*, 2020; Breznau *et al.*, 2022; Gould *et al.*, 2023; Huntington-Klein *et al.*, 2021; Silberzahn *et al.*, 2018). For example, in Silberzahn *et al.* (2018), 29 teams that included 61 analysts used the same data set to address the same research question: Are soccer referees more likely to give red cards to dark-colored players than to light-colored players? The final conclusions varied widely among the teams. In total, 20 teams (69%) found a statistically significant positive effect, and 9 teams (31%) did not observe a significant relationship. A closer look at the post-analysis survey reveals that researchers made vastly different assumptions, for example, about which variables to use as control variables.

Building on the two key theses, a consistent narrative emerges: as-

sumptions form the foundation of causal inference, and their credibility ultimately determines the validity of causal conclusions. However, in practice, researchers often make assumptions implicitly and subjectively to derive conclusions, sometimes without sufficient scrutiny. This monograph aims to deepen our understanding of assumptions. With better understanding, researchers can apply causal inference methods more rigorously and enhance the robustness of their findings. In addition, I aim to provide a systematic framework for marketing researchers to critically evaluate and approach assumptions (see Chapter 6).

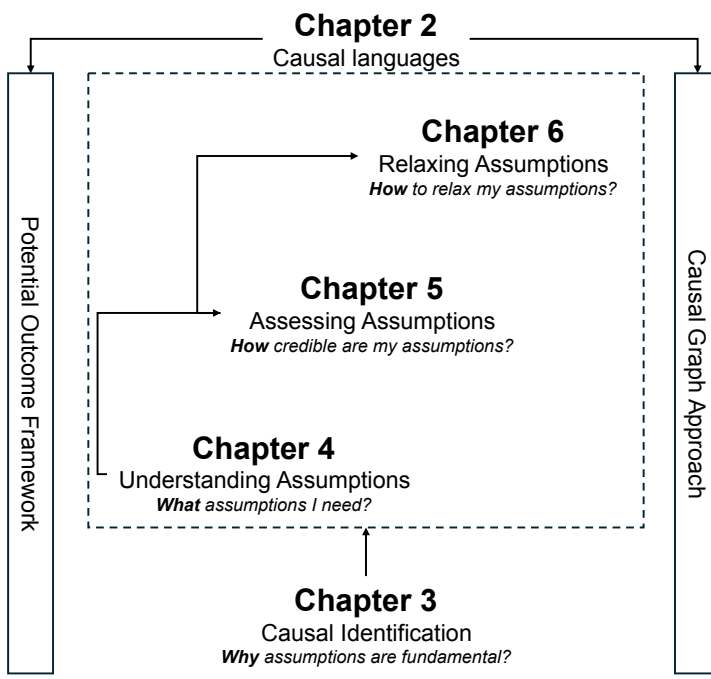
1.2 Structure of the Monograph

This monograph adopts a first-principles approach to examining assumptions in causal inference, breaking down complex problems into fundamental building blocks. To achieve this, Chapter 2 introduces the causal languages that will be used throughout the discussion, specifically the potential outcome framework (Rubin, 1974) and the causal graph perspective (Pearl, 2009). Chapter 3 establishes the centrality of assumptions in causal inference by systematically exploring causal identification and presenting three general identification strategies that underpin most research designs. Chapter 4 deepens this discussion by focusing on understanding assumptions, drawing from the philosophy of science, and illustrating their role in commonly used empirical designs such as matching, difference-in-differences (DID), and regression discontinuity design (RDD). Building on this foundation, Chapter 5 shifts to the assessment of assumptions, introducing key methods such as placebo (falsification) tests, consistency tests, and sensitivity analysis to evaluate the credibility of assumptions in empirical research. Finally, Chapter 6 presents a systematic framework for dealing with assumptions, with a particular emphasis on relaxing unrealistic assumptions to enhance the credibility of causal inference.

The structure of this monograph follows a logical progression from the establishment of foundational concepts to the application of them in practice, as illustrated in Figure 1.1. Chapter 2 provides the conceptual foundation, equipping researchers with the essential causal languages needed for later discussions. Chapter 3 addresses why assumptions are

fundamental to the credibility of casual inference. This naturally leads to Chapter 4, which focuses on understanding assumptions, ensuring that researchers can critically examine their role in different research designs. The insights of Chapter 4 lay the foundation for Chapter 5, which shifts from understanding to assessing assumptions, providing practical tools to test their credibility. Finally, Chapter 6 extends this discussion by discussing what to do when assumptions are too restrictive or unrealistic. Together, these chapters present a holistic framework that enables researchers to systematically understand, assess, and adapt assumptions, ultimately strengthening the credibility of causal inference in empirical research.

Figure 1.1: The Structure of the Monograph



2

A Brief Review of Causal Languages

Empirical research in marketing often uses statistics to study the association between a particular marketing action and a particular outcome. As one of the long-lasting themes in marketing, researchers have explored the relationship between advertising and sales since the beginning (e.g., Helmer and Johansson, 1977; Rao, 1972). However, the ultimate goal is often more ambitious; typically, marketing researchers want to establish a causal relationship, e.g., how much advertising increases sales. This is obvious, even in the early days of advertising research (e.g., Eskin and Baron, 1977; Schultz and Wittink, 1976). It is well-known that association does not imply causation. An observed positive association between advertising and sales does not automatically imply that advertising increases sales. Therefore, researchers are forever searching for the answer to this question: Under what conditions can an observed association be interpreted as a causal effect? This question is logically precluded by the key question:

What do we mean precisely when we talk about causal effects?

However, this question was not formally discussed in statistics until the late 1970s. The dominating doctrine was "statistics is about association and association is not causation." Empirical researchers who

wanted to learn causality from data often had to resort to informal reasoning. The pivotal turn occurred towards the end of the century, with two prominent paradigms of causality independently emerging. One of the paradigms, the potential outcome framework (a.k.a. Rubin causal model), was laid out by Rubin in his seminal papers (Rubin, 1974; Rubin, 1978). The Rubin causal model builds on the idea of missing individual potential outcomes and emphasizes the central role of randomization to obtain the average treatment effect. Later, Judea Pearl proposed the causal graph approach that used graph theory as the primary language of causality (Pearl, 2009). The paradigm-shifting work represents causation with a new mathematical tool. Moreover, the "do-calculus" lays out the exact mathematical conditions under which causality can be learned from data.

In this chapter, I will briefly review the two paradigms with a focus on their fundamental principles. More importantly, the review covers the causal languages that will be used in subsequent chapters. For the Rubin causal model, I focus on the fundamental problem of causal inference and its remedy based on randomization. For the Pearlian causal model, I discuss the basics of the causal graph and the foundation of "do-calculus."

2.1 The Potential Outcome Framework

Before the formal discussion, I define some notations that will be used repeatedly in the monograph. Suppose that we care about the effect of a variable D on an outcome Y. D is often referred to as "treatment" in the potential outcome framework and empirical research.

In the potential outcome framework, several fundamental concepts are essential. A **unit** (denoted as i) refers to the entity (such as a person, place, or object) upon which a treatment is applied. A **treatment** (denoted as D_i) is a variable whose causal effect on an outcome Y we want to obtain. The simplest treatment is a binary indicator of two causal states, where $D_i = 1$ if the unit i receives the treatment and $D_i = 0$ if it does not (the control condition). The potential outcomes are the outcomes under different causal states. For example, $Y_i(1)$ is the potential outcome if the unit i receives treatment and $Y_i(0)$ is the

potential outcome if it does not. The **causal effect** of a treatment is often defined as the difference in the potential outcomes. For the unit i possibly in two causal states, the **individual treatment effect** is often expressed as $\tau_i = Y_i(1) - Y_i(0)$.

2.1.1 The Fundamental Problem of Causal Inference

One of the core challenges in causal inference is known as the **fundamental problem of causal inference**. This problem arises because, for any given unit i, we can only observe the outcome under one condition, treatment or control, but not both. Specifically, we can observe either $Y_i(1)$, the potential outcome when unit i receives the treatment, or $Y_i(0)$, the potential outcome when unit i does not, but never both at the same time. This creates an inherent limitation because the individual treatment effect can never be learned for any individual unit.

The Fundamental Problem of Causal Inference

For a unit, the individual treatment effect can never be learned because only one potential outcome is observed for the unit.

To illustrate this, imagine a scenario where we evaluate the effect of an online advertisement on the purchasing behavior of consumers. For a given consumer i, we are interested in her spending amount if he/she is exposed to the ad ($Y_i(1)$) and if he/she is not exposed to the ad ($Y_i(0)$). However, we can only observe one of these outcomes because the consumer is either exposed to the ad or not. This leaves the other outcome, the *counterfactual or potential outcome*, unobserved. This leads to the fundamental problem of causal inference.

2.1.2 Resolving the Fundamental Problem

One possible solution to the fundamental problem is to assume homogeneity among individuals or across time. The assumption of homogeneity suggests that the potential outcomes $Y_i(1)$ and $Y_i(0)$ are the same for all individuals (homogeneity of units), or that the effect of a treatment on an individual is consistent over time (homogeneity of time). Under the homogneity of units, we can learn the individual treatment effect

by comparing different individuals. Similarly, under time homogeneity, a before-and-after comparison for the same individual also identifies the individual treatment effect [1]. However, the homogeneity assumption is often unrealistic in social sciences, where individuals are inherently heterogeneous, and their responses to treatments can vary widely based on environmental factors that are constantly changing.

A more practical approach is to focus on estimating the average treatment effect (ATE) rather than the individual treatment effect. The ATE is defined as the **expected** difference in potential outcomes across the finite population:

$$\tau_{\text{ATE}} = \text{E}[Y_i(1) - Y_i(0)] = \text{E}[Y_i(1)] - \text{E}[Y_i(0)] \tag{2.1}$$

This approach shifts the focus from individual causal effects to population-level effects by comparing the average outcomes for treated and untreated groups.

Suppose that we collect data on D and Y. From the data, we can learn about two conditional expectations $\text{E}[Y_i(1) \mid D_i = 1]$ and $\text{E}[Y_i(1) \mid D_i = 0]$. However, to learn the ATE, we need unconditional expectations $\text{E}[Y_i(1)]$ and $\text{E}[Y_i(0)]$. Therefore, from the data alone, we cannot know the unconditional expectations and therefore the ATE. The question remains as to how we can know the ATE from the data on D and Y. It turns out the sufficient conditions are:

$$\begin{cases} \text{E}[Y_i(1)] = \text{E}[Y_i(1) \mid D_i = 1] \\ \text{E}[Y_i(0)] = \text{E}[Y_i(0) \mid D_i = 0] \end{cases} \tag{2.2}$$

The conditions imply that the potential outcomes $Y_i(1)$ and $Y_i(0)$ are independent of the treatment assignment D_i. Alternatively, the treatment is said to be **unconfounded** or **strongly ignorable**. The conditions allow us to learn the ATE by comparing the average observed outcomes for the treatment group and the control group.

$$\tau_{\text{ATE}} = \text{E}[Y_i(1) - Y_i(0)] = \text{E}[Y_i(1)|D_i = 1] - \text{E}[Y_i(0)|D_i = 0] \tag{2.3}$$

[1] The homogeneity assumptions are often credible in natural sciences such as physics (e.g., two carbon atoms are the same) or chemistry (e.g., chemicals always stay the same until another compound is added).

In practice, how can we guarantee the condition? The most powerful method to create unconfounded treatments is through **randomization**, dating back to Ronald Fisher (Hall, 2007) and formalized by Rubin (1974) in the potential outcome framework. In a randomized experiment, units are randomly assigned to treatment ($D_i = 1$) or control ($D_i = 0$) groups, which ensures that the assignment of treatment is independent of the potential outcomes, satisfying the conditions in Equation 2.2.

2.2 The Causal Graph Approach

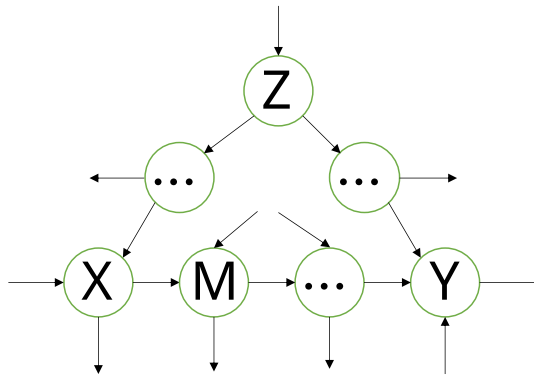
In this section, I will briefly review the basics of causal graphs and present the idea of `d`-separation using 3 basic motifs of causal graphs. I focus on the idea of "`do`-calculus", which helps researchers determine if and when a causal effect can be known from a causal graph. With the conditions from applying the "`do`-calculus," we can know the "leap of faith" needed to heist the association to causation. Therefore, the causal graph approach precisely distinguishes association and causation using the mathematical language of graph theory.

2.2.1 The Basics of Causal Graphs

A causal graph or Directed Acyclic Graph (DAG) is a graph structure that represents causal relationships between variables through directed edges. In a DAG, the nodes represent variables, and the edges (or arrows) represent the direction of causality, indicating that changes in one variable can affect another. Importantly, a DAG is **acyclic**, meaning that it does not contain any cycles or loops - no variable can cause itself directly or indirectly by following a sequence of edges [2]. A key feature of a DAG is that it represents causal flow: a variable D is said to cause an outcome Y if an **exogenous** change in D can lead to a change in Y. Figure 2.1 shows an exemplary DAG. In the DAG, $X \to M$, and X is called a *parent* of M and M a *child* of X. By the same logic, Z is called an *ancestor* of X and Y a *descendant* of X.

[2]This is to ensure the non-identification of causal effects from simultaneity, as discussed in economics literature (e.g., Manski, 1993)

Figure 2.1: An Exemplary DAG



Three fundamental assumptions are required for a DAG to serve as a valid causal model. These assumptions are comparable to the SUTVA in the potential outcome framework.

Assumption 2.1 (Three Axioms for Causal Graphs)**.**

Causal Link: *Every parent is a direct cause of all its children.*

Local Markov: *Given its parents, a node is independent of all its non-descendants.*

Faithfulness: *The correlation patterns observed in the data are always implied by the causal Markov property.*

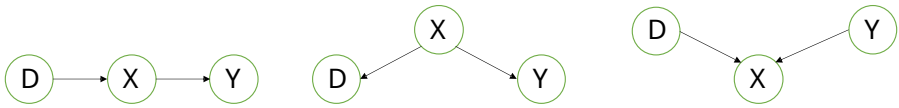
The causal link assumption ensures the existence of causal relationships, and therefore DAGs. The local Markov assumption allows for conditional independence of a node from its non-descendants, simplifying the causal inference process by enabling a local focus on a node and its direct causes. The faithfulness assumption ensures internal consistency that any observed correlation reflects true causal dependencies, ruling out the possibility of spurious correlations arising from pure statistical coincidences.

2.2.2 Studying three motifs in causal graphs

In causal analysis using DAGs, the complexity of relationships between variables can often be overwhelming. To manage this complexity, we focus on three basic motifs - repeated structures in graphs, namely the

chain, the **fork**, and the **collider** (see Figure 2.2 for their DAGs). These motifs provide simple, repeatable patterns that allow us to analyze more complex causal graphs through smaller, understandable components. By understanding these fundamental structures, we can build the foundation of how to infer causal relationships between variables in larger systems.

Figure 2.2: DAGs of Three Motifs



(a) Chain (b) Fork (c) Collision

As seen in Figure 2.2, a **chain** $D \to X \to Y$ represents a direct causal flow from one variable to another through an intermediate variable. In this structure, X has a causal influence on Y, but only through the mediator X. The **fork** motif takes the form $D \leftarrow X \to Y$, where X is a common cause of both D and Y. A **collider** motif occurs when two variables share a common effect, represented as $D \to X \leftarrow Y$.

The reason we focus on these three motifs is that they capture the essential ways in which variables can be connected in a causal system. Each motif provides a distinct type of relationship between variables: causal chains, common causes, and common effects. By understanding these basic motifs, we can decompose more complex causal structures and apply rules such as **d-separation**[3] and **do-calculus** to determine which variables are dependent or independent, given a certain set of conditions. Using the DAGs of the three motifs, one can easily observe the association and causation between D and Y, as seen in Table 2.1.

[3]D-separation is about finding a set of variables that can block all the paths between D and Y, so they are "separated" or independent. For example, in a chain, if we condition on X, the causal path from D to Y is blocked, meaning that changes in D cannot propagate to Y without going through X.

Table 2.1: Association vs. Causation in the Three Motifs

Motif	Association $\langle D, Y \rangle$	Causation $\langle D, Y \rangle$
Chain $(D \to X \to Y)$	> 0	> 0
Fork $(D \leftarrow X \to Y)$	> 0	$= 0$
Collider $(D \to X \leftarrow Y)$	$= 0$	$= 0$

2.2.3 Do-calculus: When Association = Causation?

Do-calculus is a set of rules developed to formalize the manipulation of variables in causal models represented by DAGs. The central concept of do-calculus revolves around the `do-operator`, which represents external interventions or manipulations. When we apply the `do-operator` on a variable, we sever all the incoming arrows into that variable, effectively simulating an intervention where the value of the variable is fixed externally, rather than determined by its natural causes. This is the central idea in the potential outcome framework, known as "no causation without manipulation" (Holland, 1986). The causal effect of D on the outcome Y expressed with the `do-operator` is $P(Y|\texttt{do}(D) = d)$. By this formulation, the ATE can be expressed by $P(Y|\texttt{do}(D) = 1) - P(Y|\texttt{do}(D) = 0)$.

The causal effect defined by `do-operator` highlights the difference between causation and association. The former is expressed as $P(Y|\texttt{do}(D) = d)$, while the latter is expressed as $P(Y|D = d)$. The distinction between them is critical in causal inference. $P(Y|D = d)$ represents observational data, while $P(Y|\texttt{do}(D) = d)$ simulates the effect of an intervention on D. The central question in causal inference naturally follows:

$$\text{Under what conditions } P(Y|\texttt{do}(D) = d) = P(Y|D = d) \text{ ?}$$

To this end, Pearl (2009) invented the mathematical tool of `do-calculus`, which consists of three rules that allow the transformation of expressions containing the `do-operator` into conditional probabilities, helping to evaluate causal effects even in the absence of randomized experiments. A conceptual recap of the three rules is as below [4]:

[4]For a detailed discussion of the three rules in graph theory terms, see Appendix

- **Rule 1**: Remove incoming arrows to the intervened variable when it is independent of other variables, given some set of conditions.
- **Rule 2**: If there are multiple interventions, one of them can be removed if the intervened variable is independent of the others, given certain conditions.
- **Rule 3**: You can remove interventions on certain variables if they do not influence the desired outcome, given appropriate conditioning on other variables.

Next, I will use the fork ($D \leftarrow X \rightarrow Y$) as an example to show how we can apply the rules of `do-calculus` to transform $P(Y|\texttt{do}(D) = d)$ into conditional probabilities. Equation 2.4 shows the process of transforming the causal effect of D on Y or $P(Y|\texttt{do}(D) = d)$ into conditional probabilities based on the DAG.

$$\begin{aligned} P(Y|\texttt{do}(D) = d) &= \sum_x P(Y \mid \texttt{do}(D) = d, X = x) P(X = x \mid \texttt{do}(D) = d) \\ &= \sum_x P(Y \mid D = d, X = x) P(X = x \mid \texttt{do}(D) = d) \\ &= \sum_x P(Y \mid D = d, X = x) P(X = x) \end{aligned} \tag{2.4}$$

The first equality comes from the fact that in a fork X is the common parent of D and Y. The second equality is obtained by applying the first rule of `do-calculus` based on the "backdoor criterion." The third equality follows directly from the causal link assumption (Assumption 2.1). In the fork, X is a parent of D, so manipulating D would not influence X.

2.3 A Comparison of Two Frameworks

The potential outcome framework and the causal graph approach to causality both aim to understand and infer causal relationships, but do so differently. The key converging points of the two frameworks are the idea of manipulations and the concept of "counterfactuals." In the potential outcome framework, the causal effect is defined in terms of potential (counterfactual) outcomes under different treatment

A

conditions, with the counterfactual being the unobserved potential outcome. Similarly, in the DAG approach, the concept of manipulation is central and is captured by do-operation ($\texttt{do}(D)$). Judea Pearl's ladder of causation places both approaches in the realm of interventions and counterfactuals, where the second rung (interventions) corresponds to the manipulation of variables, and the third rung (counterfactuals) addresses hypothetical alternate realities (Pearl and Mackenzie, 2018).

Despite shared insights, the two frameworks diverge in their strategies for achieving causal identification. The potential outcome framework emphasizes randomization as the gold standard for establishing causality, where randomization ensures that the potential outcomes are independent of the assignment of treatment. In contrast, the causal graph approach does not rely on randomization, but rather uses do-calculus to transform do-operations into conditional probabilities. By exploiting the structure of the DAG, the goal of do-calculus is to find conditions under which the effect of an intervention can be expressed in terms of observable data, without the need for randomization. This difference highlights a key philosophical distinction: while the potential outcome framework sees randomization as the primary tool for causal inference, the causal graph approach seeks to uncover causal relationships from observational data through mathematical transformations.

3

Causal Identification and Assumptions

Suppose a researcher collects data on sales and prices and runs a regression $\text{Sales} = \beta\text{Price} + \varepsilon$ to learn β. Taking the conditional expectation of Price on the equation, we have $\text{E}(\text{Sales}|\text{Price}) = \beta\text{Price} + \text{E}(\varepsilon|\text{Price})$. Assume that the researcher has **perfect data** on Prices and Sales, with an infinite sample size and perfect measurements. With these data, the researcher can learn the conditional expectation $\text{E}(\text{Sales}|\text{Price})$ directly from data. To know the price effect β, we still need $\text{E}(\varepsilon|\text{Price})$. However, we cannot learn $\text{E}(\varepsilon|\text{Price})$ from the data as ε is not observed.

This example illustrates that we may not be able to learn the wanted parameters or estimands from even perfect data. If not identified, multiple parameter values could be consistent with the same observed data, leading to ambiguity. This is exactly why identification is so fundamental in causal inference. However, key questions are yet to be answered. For example, what is the nature of the identification process? What are the basic building blocks of identification? How do assumptions play a role in identification? Can we synthesize some general processes or identification strategies from a plethora of different causal inference methods? In this chapter, I will address these questions. The following proposition (3.1) summarizes the main thesis of this chapter.

The Essential Role of Identification in Causal Inference

The rigor and credibility of causal inference methods come from the formal treatment of identification, where the assumptions that are required to learn the causal estimand are formally derived and transparently communicated.

3.1 Introduction to Identification

Assumptions are the cornerstone of the identification process, as they define the conditions under which a model parameter, such as a treatment effect, can be uniquely inferred from observable data. At its core, identification is about determining whether the available data, when combined with a set of structural assumptions, are sufficient to recover the estimand of interest. However, the process does not deal with practical concerns like sample size limitations or measurement errors; instead, it operates under the idealized premise that the true joint distribution of variables is known. This abstraction allows researchers to focus on the logical structure of inference, ensuring that causal parameters can be learned, in principle, before estimation is attempted. In this sense, assumptions do not just support identification; they enable it by providing the necessary framework to connect observed data with causal relationships.

Critically, assumptions resolve the fundamental issue of non-identifiability, where multiple parameter values could be consistent with the same observed data, leading to ambiguity. Without a well-defined set of assumptions, it is impossible to distinguish between competing explanations, rendering estimation efforts meaningless.

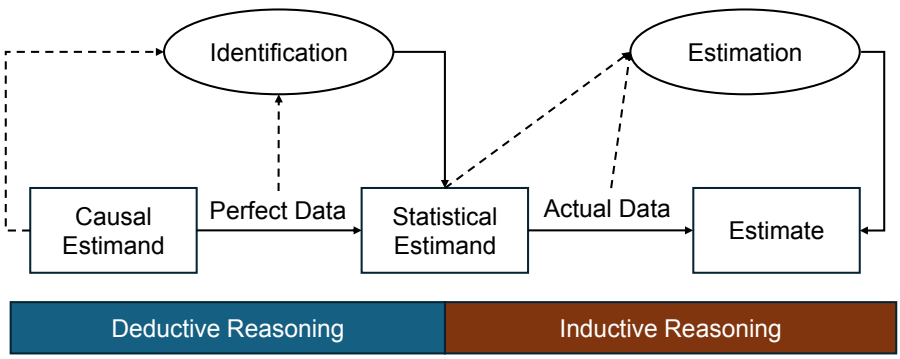
3.1.1 The Basic Idea of Identification

The basic idea of identification is to determine whether a model or a parameter (such as a treatment effect) can be inferred unambiguously from the observable distributions of variables in the data (Lewbel, 2019). There are two things that need further clarification here. First, the definition assumes that we already have an objective in mind (a

parameter or an estimand). Second, we "distill" data into observable distributions of variables, that is, we assume that we can learn the true (joint) distribution of the variables from the data. This therefore assumes away the typical statistical problems such as sample sizes and measurement errors. In other words, it is assumed that the data are assumed to be "perfect" in the identification process.

The definition of identification implies that identification naturally proceeds with the estimation (Lewbel, 2019). Before we can apply any estimation technique to infer parameters or causal effects from data, we must first establish that these parameters can be learned from the data ("identified"). In other words, identification is the process of determining whether the parameters of interest can be recovered from the available data given a set of assumptions. Without identification, estimation efforts are futile because multiple parameter values could produce the same observable data, leading to ambiguity and inconsistency. This relationship between identification and estimation is illustrated in the flow chart (Figure 3.1).

Figure 3.1: The Identification-Estimation Flowchart



Lastly, I will discuss identification from a mathematical logic point of view. The central idea of identification is the logical reasoning of how an estimand can be learned from data. In the identification process, the most important part is to **find a sufficient set of assumptions** [1]

[1]These assumptions are also referred to as "data-generating processes" in eco-

that allow us to learn the estimand given the "perfect data." Therefore, the identification process can be viewed as an **axiomatization process**, which defines a system or framework by specifying a set of basic assumptions or principles, known as axioms, from which other truths or conclusions can be logically derived.

In various fields such as mathematics, logic, and economics, the axiomatization process involves distilling complex systems into their most essential assumptions or principles, which serve as the starting point for all further analysis. For example, in utility theory in economics, one of the axioms known as "transitivity" in the Von Neumann–Morgenstern utility theorem (Von Neumann and Morgenstern, 2007) states that preferences are consistent across all three options. In this stage of the identification-estimation flow, the thinking process of deductive reasoning is applied, where one starts with a set of assumptions and works toward the conclusion that the causal estimand is identified.

Identification as An Axiomatization Process

Identification is an axiomatization process, where one finds a set of sufficient assumptions and uses deductive reasoning to conclude that the estimand is identified.

Although I have established the basic elements of identification, in practice, this concept can be confusing. This is partially attributed to the way that statisticians and econometricians give different meanings to identification. Furthermore, within economics, many terms related to identification appear in the literature (Lewbel, 2019). For clarity, I will focus on three types of identification, namely causal identification, point identification, and set identification. Point identification is the most common form of identification in empirical research. The identification of sets, although less commonly adopted, helps illustrate the relationship between assumptions and conclusions. Lastly, I will discuss causal identification in detail.

nomics. For example, see (Davidson and MacKinnon, 1993)

3.1.2 Point Identification

I will first discuss the **point identification** and in the meantime illuminate some related concepts in identification. The treatment effect (ATE) will be used as the exemplary estimand, but the idea applies to other estimands. The discussion here follows closely those in Lewbel (2019) and Matzkin (2007). The concept of point identification is the process of determining whether the parameter of interest or the focal estimand, denoted $\theta \in \Theta$, can be uniquely determined from observable data, represented by Φ. Θ is the set of all admissible values of θ, and Φ captures the information in the data, usually in the form of the joint distribution of variables. Without further assumptions, there are often **observationally equivalent** parameters that produce the same joint distribution or Φ. **Observational equivalence** occurs when two different parameter values, θ and $\tilde{\theta}$, imply the same Φ. Point identification is achieved if there are no two different values θ and $\tilde{\theta}$ that are observationally equivalent.

To achieve point identification, we often need to make a set of **sufficient assumptions**. Assumptions can take different specific formats. They can be *conceptual* (e.g., consumers self-select into the exposure to advertising), *mathematical* (e.g., the model is linear), or *statistical* (e.g., the error term follows a normal distribution with mean zero). Let $s \in S$ represent a set of assumptions in all possible sets of assumptions S. Given the set of assumptions s and the estimand θ, we have a structure (or a tuple) $t = \langle s, \theta \rangle$ that can produce the data Φ [2].

The key challenge in identification is to ensure that the true parameter θ_0 is uniquely quantified under the set of assumptions s. This means that no other parameter $\tilde{\theta}$ can generate the same observable data Φ, given assumptions s. Or equivalently, different values of θ must imply different observable implications for the data. In general, in point identification, we try to find a set of assumptions to ensure that there is no observational equivalence between the values of distinct parameters [3].

[2] We call a structure t admissible if it can produce data patterns that are consistent with Φ. Here, we focus only on admissible structures

[3] Note that the set of assumptions needs not to be unique. We may have multiple

To illustrate the concept of identification and observational equivalence, consider the example of treatment effects. Assume that we want to identify the ATE in the context of a binary treatment. The outcome Y is observed for people assigned to treatment $D = 1$ or control $D = 0$, and the objective is to identify θ, the ATE, defined as: $\theta = E[Y(1) - Y(0)]$, where $Y(1)$ and $Y(0)$ represent the potential outcomes under treatment and control, respectively. Without loss of generality, assume that the true model is $Y = \theta D + \varepsilon$, and therefore the ATE is equal to θ. From the model, the observed differences in the expected outcomes is,

$$\underbrace{\mathrm{E}\left(Y\left(1\right) \mid D = 1\right) - \mathrm{E}\left(Y\left(0\right) \mid D = 0\right)}_{\text{Observed Differences}} = \underbrace{\theta}_{\text{True ATE}} + \mathrm{E}\left(\varepsilon \mid D = 1\right) - \mathrm{E}\left(\varepsilon \mid D = 0\right) \tag{3.1}$$

In this setup, the information from the data Φ consists of the conditional expectations $\mathrm{E}\left(Y\left(1\right) \mid D = 1\right)$ and $\mathrm{E}\left(Y\left(0\right) \mid D = 0\right)$. From Equation 3.1, when the assumption of *unconfoundedness* holds, i.e., the treatment assignment D is independent of the potential outcomes $Y(1)$ and $Y(0)$, the ATE is point identified. However, without the assumption, we have no knowledge about $\mathrm{E}\left(\varepsilon \mid D = 1\right)$ and $\mathrm{E}\left(\varepsilon \mid D = 0\right)$. Different values of θ's can produce the same conditional expectations in Φ. Therefore, the unconfoundedness assumption $s \equiv \varepsilon \perp D$ allows θ to be point identified.

3.1.3 Set Identification

In contrast to point identification, where the parameter θ is uniquely determined from the observable data, **set identification** occurs when the data only allow us to pin down a range or set of possible values for θ, rather than a unique value. In this case, the true parameter θ_0 lies within a set, called the *identified set*, denoted as Θ_I. Formally, a parameter is *set identified* if given the information in the data Φ, there is a set of values for θ, denoted by $\Theta_I(\Phi)$, such that all elements in the

sets of assumptions that allow for point identification. In the price effect example, instead of assuming $\mathrm{E}(\varepsilon \mid \text{Price}) = 0$, we can let it equal to any other specific value, which also achieves point identification for β.

set are consistent with the observed data. Mathematically, this can be expressed as $\theta_0 \in \Theta_I(\Phi)$ (Chesher and Rosen, 2017).

Set identification arises when the assumptions are insufficient for point identification, meaning that different values of θ could generate the same observable pattern in data Φ. In such cases, instead of a unique parameter estimate, we obtain a range of plausible values for θ, which reflects the partial nature of the identification (Tamer, 2010). Key to set identification is recognizing that, even though a parameter may not be precisely pinpointed, valuable information can still be derived by narrowing the possibilities down to a credible range. As highlighted by Charles Manski (Manski, 2003), various types of restrictions can help narrow the identified set, even when point identification is not feasible.

The question remains why set identification deserves our attention. The most important reason is the relationship between the strength of assumptions and the credibility of our conclusions [4]. Proposition 1.1 implies that empirical researchers always need to make assumptions, but how assumptions influence our conclusions? This question is best answered by the Law of Decreasing Credibility (see p.1 of Manski, 2003). This principle suggests that empirical researchers encounter a trade-off when choosing which assumptions to adopt: stronger assumptions can lead to more robust inferences but may come at the cost of reduced credibility.

The Law of Decreasing Credibility

The credibility of the inference decreases with the strength of the assumptions maintained.

To illustrate the concept of set identification, let us revisit the example of estimating the ATE from the previous section. Under the assumption of unconfoundedness, the ATE can be point identified, as shown earlier. However, if unconfoundedness does not hold, point identification may fail and instead the ATE may be non-identified. In this case, we may only determine that the ATE lies within an interval,

[4] Set identification also arises due to the incompleteness of the underlying models, for example, the multiple equilibria of game-theoretic models (Tamer, 2003).

rather than finding its exact value.

Charles Manski's frameworks for *monotone treatment response (MTR)* (Manski, 1997) and *monotone treatment selection (MTS)* (Manski and Pepper, 2000) provide examples of how ATE can be set identified. Under the MTR assumption, the individual treatment effect $\tau_i = Y_i(1) - Y_i(0)$ is assumed to be non-negative (or non-positive) for all individuals, implying that treatment always improves (or worsens) outcomes. This restriction helps to narrow the identified set for the ATE, even when point identification is not possible. Similarly, under MTS, it is assumed that individuals who receive treatment have better (or worse) potential outcomes than those who do not, based on an observable characteristic. These monotonicity assumptions allow us to identify bounds on the ATE, such that $\theta_0 \in [\theta_{\text{lower}}, \theta_{\text{upper}}]$, where θ_{lower} and θ_{upper} represent the lower and upper bounds. Thus, even in the absence of point identification, the ATE can be meaningfully constrained within an interval, providing partial identification of the causal effect.

3.2 General Causal Identification Strategies

As illustrated in 3.1, causal identification is a crucial aspect of causal inference, which aims to determine the conditions under which a causal effect can be inferred from observed data. As the identification problems discussed in the previous section, causal identification has the same problem structure. In this section, I will extensively use causal graph languages to discuss the problem of causal identification and present three general causal identification strategies. These general strategies are foundational to understand specific research designs. For example, the matching method (Rubin, 1973a; Rubin, 1973b) is essentially the conditioning strategy.

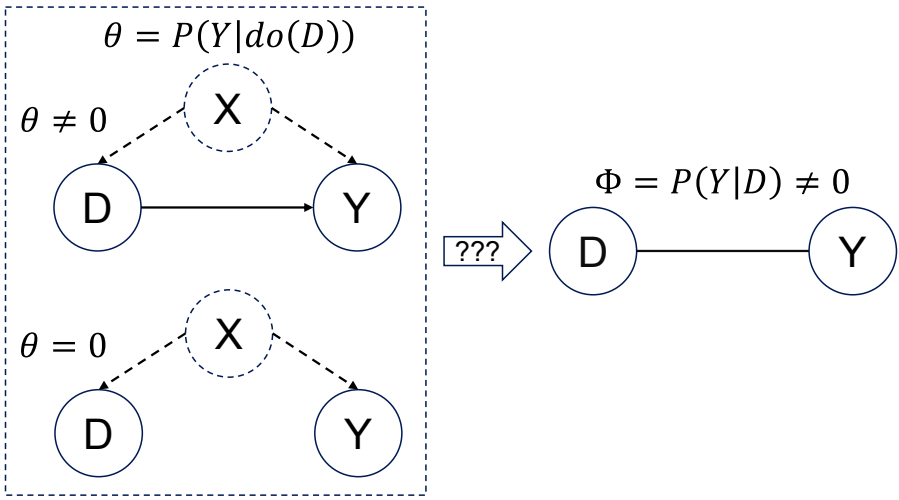
I will first formalize the causal identification problem. In practice, empirical researchers often care about the causal effect of a treatment D on an outcome Y, with $\theta = P(Y \mid \texttt{do}(D))$. The data collected on D and Y give us information on the conditional probability, with $\Phi = P(Y \mid D)$. However, the conditional probability directly learned from the data is often unequal to the causal effect - "association is not causation." In

this case, multiple causal effects θ can produce the same patterns in Φ, as illustrated in Figure 3.2. In empirical research, the causal effect of D on Y is often confounded by the confounders X, and researchers only observe a partial set of X. Researchers then need to devise different approaches to solve this issue. This becomes the fundamental problem in empirical research, which I term as the *core problem of causal inference*.

The Core Problem of Causal Inference

The effect of the cause variable D on the outcome Y is confounded by X, which in the best case is partially observed. Researchers must devise different approaches to solve this problem.

Figure 3.2: The DAG for the Core Problem



To resolve the issue, researchers must go through the due identification processes by finding a set of sufficient assumptions that guarantee that the causal estimand θ can be learned from the data Φ. Given specific empirical settings, researchers can adopt different research designs, strategies, and frameworks that researchers use to identify and estimate causal effects from data (Angrist, 2022). For clarification, I view the terms "research designs", "empirical strategies" and "identification

strategies" as interchangeable. For consistency, I will use "identification strategies" throughout the monograph. In this chapter, I will focus on three general identification strategies underlying the specific stylized research designs. For each strategy, I will discuss its assumptions and how identification is achieved given the assumptions.

3.2.1 Conditioning Strategy

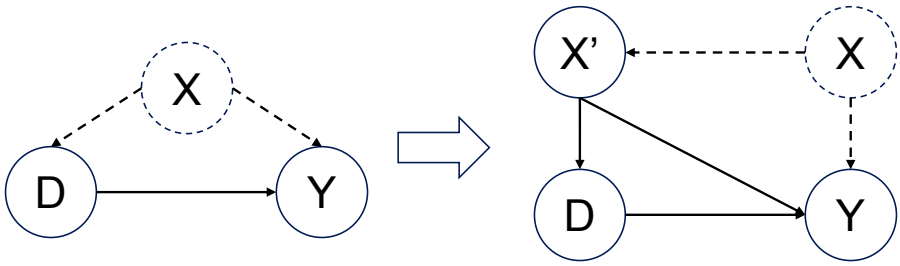
The first general strategy to solve the core problem of causal inference is based on the "backdoor criterion" (see p.79 of Pearl, 2009) - **conditioning strategy**. As seen from the DAG in Figure 3.3, X form backdoor paths between D and Y, creating spurious correlations between them. The backdoor paths are non-causal paths ways between the treatment D and the outcome Y. The conditioning strategy therefore makes an assumption that we observe a set of variables X' that block all the backdoor paths [5]. With this assumption, we can isolate the causal effect by ensuring that any association between treatment and outcome is not confounded by other variables. This conditioning strategy makes an **exhaustiveness assumption** that all relevant confounders are observed and can be accounted for by including them in the analysis.

The conditioning strategy is adopted widely in specific research designs. One common research design based on the conditioning strategy is *matching*, where treated and control units are paired based on their similarity on observed covariates, thus mimicking random assignment. For matching, the key assumption is the conditional exogeneity assumption that the treatment D is independent of potential outcomes $Y(D)$, conditional on the set of observables X. Another example is *regression analysis*, where confounders are included as control variables to estimate the causal effect of the treatment, with the key assumption that the error term of the regression is exogenous given the control variables. The propensity score matching, stratification, and control variables in regression models are all applications of the conditioning strategy, aimed at reducing bias due to confounding.

To see how the exhaustiveness assumption identifies the effect $P(Y \mid$

[5]Note that X' can be different from X in that we can condition on the variables that are descendants of X and parents of D to block the backdoor paths.

Figure 3.3: The DAG for the Conditioning Strategy



$\texttt{do}(D)$). We can apply the backdoor criterion, which states that if a set of variables X' blocks all backdoor paths (non-causal paths) from the treatment D to the outcome Y, then the causal effect of D on Y can be identified by conditioning on X'. Formally, if X' satisfies the backdoor criterion, then the causal effect $P(Y \mid \texttt{do}(D))$ can be estimated by adjusting for X' using the standard conditional probability $P(Y \mid D, X')$. This ensures that the non-causal pathways between X and Y are properly controlled, allowing for valid causal identification.

$$P(Y \mid \texttt{do}(D)) = \sum_{x} P(Y \mid D, X' = x)P(X' = x) \tag{3.2}$$

The same result can be obtained by applying the potential outcome framework, assuming D is binary. Note that the data Φ now include D, Y and X'. From the exhaustive assumption, we have the conditional unconfoundedness that $Y(1), Y(0) \perp D \mid X'$. This implies that, conditioning on $X' = x$, we can identify the ATE:

$$\begin{aligned} \mathrm{E}(Y(1) - Y(0) \mid X' = x) = \mathrm{E}(Y(1) \mid D = 1, X' = x) \\ - \mathrm{E}(Y(1) \mid D = 0, X' = x) \end{aligned} \tag{3.3}$$

The ATE can then be constructed by the following equation:

$$\begin{aligned} \mathrm{E}(Y(1) - Y(0) \mid X' = x) = \sum_{x} (\mathrm{E}(Y(1) \mid D = 1, X' = x) \\ -\mathrm{E}(Y(1) \mid D = 0, X' = x)) \end{aligned} \tag{3.4}$$

The conditioning strategy lays the foundation for many specific methods in causal inference, such as matching, difference-in-differences

(DID) and synthetic control methods (SCM). These methods apply the conditioning strategy by controlling for confounding variables to isolate causal effects. Matching pairs up treated and control units with similar features to isolate the treatment effect (e.g., Rubin, 1973a). Difference-in-differences (DID) utilizes pre- and post-treatment periods to control for unobserved individual factors that are invariant in time (e.g., Lechner *et al.*, 2011). The synthetic control method constructs a control unit that closely resembles the treated unit in the pre-treatment period, effectively conditioning on pre-treatment characteristics (e.g., Abadie *et al.*, 2010). In each case, the conditioning strategy underlies the approach.

3.2.2 Identification by Mechanism

The **identification by mechanism** strategy, also known as the front-door criterion, allows causal identification by leveraging intermediate variables that fully mediate the effect of treatment D on the outcome Y. In this strategy, a set of exogenous variables M is used to mediate the causal effect of D on Y, thus isolating the mechanisms involved. According to the front-door criterion (see p.81 of Pearl, 2009), a set of variables M satisfies this criterion when: (i) **Exhaustiveness** - M blocks all directed paths from D to Y; (ii) **Unconfoundedness** - there are no unblocked back-door paths from D to M; and (iii) **Uniqueness** - D blocks all back-door paths from M to Y. Under these conditions, the causal effect of D on Y can be expressed as:

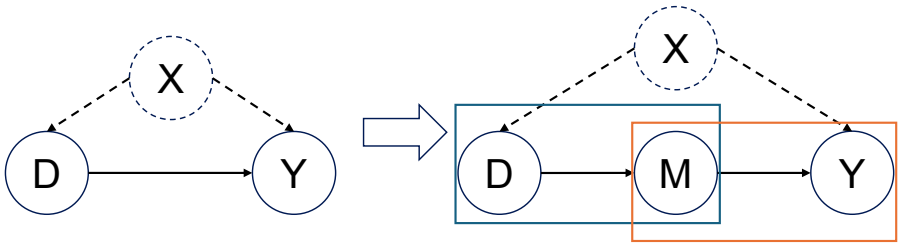
$$P(Y \mid \texttt{do}(D)) = \sum_m P(M = m \mid D) \sum_y P(Y = y \mid M = m), \qquad (3.5)$$

where the probabilities represent the observable data and can be used to estimate the causal effect through mediation. This approach relies heavily on identifying the appropriate mediating variables and ensuring that the conditions of the front-door criterion are met.

Figure 3.4 intuitively shows how the effect $P(Y \mid \texttt{do}(D))$ is identified under the three assumptions. More formally, the identification can be proved by applying the rules of `do`-calculus.

The identification results can be visualized by the DAG in Figure 3.4. Formally, we can use `do`-calculus to and the DAG to transform the

Figure 3.4: The DAG for the Identification by Mechanism



causal estimand $P(Y \mid \texttt{do}(D))$ into conditional probabilities.

$$\begin{aligned} P\left(Y \mid \texttt{do}\left(D\right)\right) &= \sum_{m} P\left(Y \mid \texttt{do}\left(D\right), M=m\right) P\left(M=m \mid \texttt{do}\left(D\right)\right) \\ &= \sum_{m} P\left(Y \mid M=m\right) P\left(M=m \mid \texttt{do}\left(D\right)\right) \\ &= \sum_{m} P\left(Y \mid M=m\right) P\left(M=m \mid D\right) \\ &= \sum_{m} P\left(M=m \mid D\right) \sum_{y} P\left(Y=y \mid M=m\right) \end{aligned} \tag{3.6}$$

The first equality is an application of the law of total probability, given M is **unconfoudned** in the DAG. The second equality is to apply the third rule of `do`-calculus, since all front-door paths between D and Y are blocked by M (**exhaustiveness**). We can then remove $\texttt{do}(D)$ from the formula. The third equality is the assumption of causal link and the fact that the $D \to M$ path is unconfounded or without any backdoor path (**uniqueness**). The last equality is a simple rearrangement of terms.

For identification by mechanism to be valid, the key assumptions must hold. These assumptions are critical because they ensure that the mediating variables M provide a reliable pathway to estimate the causal effect of D on Y. The **exhaustiveness** assumption implies that no direct paths from D to Y exist outside of those mediated by M. The **unconfoundedness** assumption ensures no spurious associations, and both $P(M \mid \texttt{do}(D))$ and $P(Y \mid \texttt{do}(M))$ are identified. The **uniqueness** assumption assumes away any other causes of Y, which could spuriously

attribute their influence on Y to M. These assumptions mirror the conditions necessary for satisfying the front-door criterion in causal graphs, and serve as the foundation for using the identification-by-mechanism strategy to estimate causal effects.

However, empirical applications of this strategy are infrequent, most likely due to the difficulty of finding an exogenous "super mechanism" in many empirical settings. Cohen and Malloy (2014) applies the **identification by mechanism** strategy to study the logrolling behaviors in congressional voting. Specifically, they study how the alumni relationship between senators influences the *quid pro quo* in bill voting. They exploit a specific mechanism - the quasi-randomly assigned seating during voting sessions - as the super mechanism. In the empirical setting, two senators can only logroll if they are seated closely, due to the lack of other ways of communication. Therefore, the realization of the social effect of being alumni fully depends on the quasi-randomly assigned seating.

3.2.3 Instrumental Variable Strategy

The instrumental variable (IV) strategy is another strategy to solve the core problem of causal inference. It has roots in the field of econometrics and originated as a solution to address endogeneity. The concept was first introduced by Philip G. Wright in his 1928 book, The Tariff on Animal and Vegetable Oils, where he used it to study the supply and demand for agricultural products. Further formalization of the IV approach came in the mid-20th century with work by early econometric work such as Haavelmo (1943) and Koopmans (1953). From then on, the IV method became a key tool in empirical research, especially in settings where randomized experiments were impractical. Since then, instrumental variables have played a critical role in causal inference in multiple disciplines, including economics, epidemiology, and social sciences.

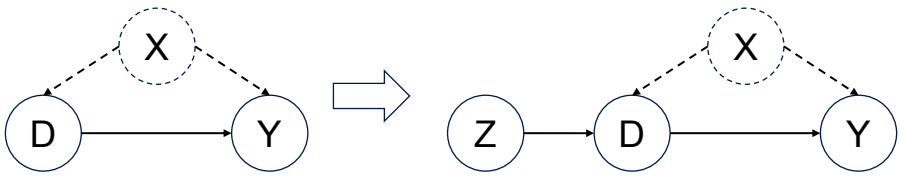
The IV strategy solves the core problem of causal inference by introducing an exogenous variable Z, known as an instrument, which affects Y only indirectly through D. In doing so, the IV strategy transforms the DAG of the core problem. Figure 3.5 shows the original DAG and

the transformed DAG by including an IV Z. The IV strategy essentially relies on an exogenous shock Z and the fully mediated causal path $Z \to Y$ (by the cause variable D). Under the transformed DAG, two causal effects $P(Y \mid \text{do}(Z) = z)$ and $P(D \mid \text{do}(Z) = z)$ can be learned directly, due to the exogeneity of the instrument Z. From the DAG, we have the following equation:

$$P(Y \mid \text{do}(Z) = z) = \sum_{d} P(Y \mid \text{do}(D) = d) P(D = d \mid \text{do}(Z) = z) \quad (3.7)$$

In the equation, $P(Y \mid \text{do}(Z) = z)$ and $P(D \mid \text{do}(Z) = z)$ are identified from data $\Phi = P(Z, D, Y)$. The unknown in the equation is $P(Y \mid \text{do}(D) = d)$, and it can be obtained by solving the system of equations.

Figure 3.5: The DAG for the Instrumental Variable Strategy



For Z to be a valid instrument, it must satisfy two main conditions: 1) **Exogeneity**: Z should not be related to any confounders (i.e., $Z \perp X$), and 2) **Exclusion**: There should be no unblocked paths from Z to Y that bypass D, meaning that Z influences Y only through D. In addition, we need two other regularity conditions as extra assumptions, namely **relevance** and **monotonicity**. The **relevance** assumption rules out the trivial conditions that satisfy the system of equations in Equation 3.7, where both sides are equal to zero, i.e., $P(Y \mid \text{do}(Z) = z) = 0$ and $P(D \mid \text{do}(Z) = z) = 0$. The **monotonicity** assumption guarantees the uniqueness of the solutions of Equation 3.7.

With these assumptions, we can compute the causal effect $P(Y \mid \text{do}(D) = d)$ by solving Equation 3.7 in terms of $P(Y \mid \text{do}(Z) = z)$ and $P(D \mid \text{do}(Z) = z)$. In case where all relationships among variables are linear, the IV strategy is often estimated using methods such as two-stage least squares, where Z serves as an instrument to estimate the endogenous variable D without confounding bias.

In the search for valid instruments, marketing researchers often look to **natural occurrences**, **policy changes**, or **institutional processes** that can provide exogenous sources of variation. For example, Li and Ching (2023) uses the "Death IV" pioneered by Azoulay *et al.* (2019) where the unexpected deaths of key figures provide variation that is exogenous to other factors influencing socially impacted outcomes, such as scientific productivity or job opportunities. Shriver *et al.* (2013) use exploit the wind speed at surfing locations in Switzerland as a source of variation in users' propensity to post content about their surfing activity on an online social network. These are examples of how natural occurrences can be a good source of IVs. In marketing, researchers also rely on policy changes, such as election cycles, as an exogenous shock to study marketing mixes, such as advertising (e.g., Moshary *et al.*, 2021). Lastly, knowledge about certain institutional processes also enables researchers to construct useful instruments. For example, Hui *et al.* (2013) use the store layout to construct a "reference path" for the actual path taken by a consumer in a retailing store.

Although the IV strategy is widely adopted in empirical work in various fields, including marketing, there are a few key challenges. A central challenge is the need for strong assumptions, namely exogeneity, exclusion, and relevance, to hold in order for the IV estimates to be valid. These assumptions are often difficult to verify empirically and in many cases cannot be statistically tested, which limits the reliability of IV estimates when assumptions are questionable. Furthermore, there is the issue of weak instruments (Andrews *et al.*, 2019): If the instrument Z has only a weak correlation with the endogenous variable D, the resulting estimates can be biased and imprecise. This problem is exacerbated when using IVs in finite samples, as weak instruments can inflate standard errors, leading to misleading inference.

In sum, it is often challenging to find suitable instruments that satisfy all the necessary conditions. Empirical researchers must often rely on theoretical arguments or external knowledge to justify their choice of instruments, which can introduce subjectivity and weaken causal claims when assumptions are debatable. In addition, the estimated causal effects with the IV strategy are often difficult to interpret. In fact, a clear and precise interpretation is not available until Imbens and Angrist

(1994), which we discuss in detail in Chapter 4.

3.3 Assumptions in Causal Identification: A Snythesis

In the previous two sections, I explore the concept of identification in the context of causal inference. Identification addresses the core problem of causal inference: distinguishing causation from correlation when the causal effect of an intervention or treatment on an outcome may be confounded by unobserved factors. The key idea is that identification is a prerequisite to estimation - before any causal effect can be consistently estimated, it must first be identifiable under the given assumptions.

In practice, researchers generally apply three broad identification strategies (or their combinations): the conditioning strategy, the identification by mechanism, and the instrumental variable strategy. Each strategy offers a distinct approach to overcoming the core problem of causal inference and isolating causal effects based on different assumptions. The conditioning strategy relies on the backdoor criterion, assuming that all relevant confounders are observed and can be conditioned upon. Identification by mechanism, or the front-door criterion, utilizes a mediator that fully transmits the causal effect from treatment to outcome, given that the mediator itself is unconfounded. Finally, the IV strategy introduces an external source of variation - an instrument - that influences the treatment but is independent of the outcome except through the treatment.

Together, these identification strategies form a foundational framework for empirical research in related fields such as economics and marketing. Each strategy is suited for different empirical settings, and the choice of strategy depends on the nature of the available data and the plausibility of the required assumptions. Ultimately, effective causal inference relies on carefully chosen assumptions, grounded in theoretical justification and practical feasibility, to ensure credible identification and, subsequently, consistent estimation of causal effects.

In general, the essential role of identification in empirical research puts assumptions in a pivotal position. As highlighted in Proposition 1.1, the credibility of conclusions is based on the credibility of assumptions. The pivotal role of assumptions in scientific inquiries is becoming

increasingly recognized by researchers. Recent studies in various fields show that scientists reach vastly different conclusions for the same research questions when given the same data (Botvinik-Nezer *et al.*, 2020; Breznau *et al.*, 2022; Gould *et al.*, 2023; Silberzahn *et al.*, 2018). The data collected in these studies reveal that researchers make different assumptions in their analysis, which in turn lead to different conclusions. In the next chapter (Chapter 4), I will focus on understanding assumptions.

4

Understanding Assumptions

Assumptions form the foundation of causal inference, which guides researchers in their efforts to establish credible causal relationships. The validity of causal claims is inherently tied to the assumptions that underpin them. In this chapter, I discuss how to understand assumptions from a broader epistemological perspective and then zoom in on assumptions of stylized research design.

From a broader epistemological perspective, assumptions are not merely technical necessities but also fundamental elements of scientific reasoning. They serve as conceptual simplifications that make complex problems tractable, allowing researchers to construct models that approximate reality. However, the use of assumptions introduces tension for researchers (Proposition 4.1). Although assumptions enable causal identification, they also impose constraints that can limit the credibility of conclusions. As highlighted in the philosophy of science, assumptions can be explicit or implicit, testable or untestable, and sufficient or necessary. Understanding both their necessity and their limitations is crucial for interpreting empirical results.

The chapter then delves into the assumptions underlying various causal inference methods, including instrumental variables, matching,

difference-in-differences (DID), regression discontinuity design (RDD), and synthetic control methods (SCM). By analyzing the sufficiency, necessity, testability, and epistemological status of these assumptions, this chapter aims to illuminate their implications for causal identification and estimation. Through this discussion, researchers can better understand the strengths and limitations of different methodological approaches and adopt a more critical and transparent approach in their causal inference studies.

4.1 The General Understanding of Assumptions

So far, I have established the importance of assumptions in causal identification. However, the foundational role of assumptions in scientific inquiry has long been recognized in the philosophy of science. Assumptions are essential in shaping theories, guiding empirical research, and interpreting findings. They help simplify complex phenomena, making systematic analysis possible. Beyond their practical use, assumptions define the scope of scientific inquiry and influence how evidence is understood. In this section, I will discuss the general understanding of assumptions from the perspective of the philosophy of science.

4.1.1 The Fundamental Role of Assumptions in Scientific Inquiry

Many philosophers of science believe that assumptions are the foundational elements of scientific inquiry, acting as conceptual scaffolding that enables scientists to develop theories, design experiments, and interpret data. As Giere (1999) notes in *Science without Laws*, assumptions provide the scaffold for simplifying complex phenomena, making them amenable to systematic study and analysis. For example, rational agents are necessary abstractions that allow economists to derive generalizable principles. In marketing, one of the fundamental assumptions is that consumers are different, which formulates the foundation of decisions of marketing mixes (Palmatier and Crecelius, 2019). Similarly, Cartwright (1983) in *How the Laws of Physics Lie* emphasizes that scientific models are built on assumptions that often deliberately oversimplify reality to focus on causal relationships, revealing mechanisms that might other-

wise remain obscured. Similarly, assumptions play a pivotal role in the investigation of causality.

However, assumptions are not merely tools of convenience; they are deeply embedded in the epistemological frameworks of science, shaping the ways knowledge is constructed, validated, and interpreted (Kuhn, 1997). In many cases, assumptions provide the foundational premises that make scientific inquiry possible by defining what is observable, measurable, and explainable within a given framework. For example, in economics, the *Homo economicus* assumption underpins the rational choice theory, whereas in statistical modeling, assumptions about independence or normality determine the validity of inferential methods. These underlying assumptions are critical because they not only guide the formulation of theories but also influence how evidence is interpreted and integrated into broader scientific understanding. As such, the credibility of scientific theories and conclusions hinges on the plausibility of their assumptions.

In general, assumptions that align with empirical evidence or established theoretical frameworks tend to bolster the reliability of a theory, while unexamined or implausible assumptions can undermine its legitimacy. Moreover, as in Cartwright (1983), the utility of assumptions in simplifying complex systems often comes at the cost of omitting key factors, raising questions about the degree to which such models accurately reflect reality. Ultimately, assumptions play a dual role in science: they are indispensable for constructing and testing theories, but their epistemological influence demands constant scrutiny to ensure that they do not compromise the credibility of scientific conclusions. Causal inference, as a scientific field, is also empowered and constrained by assumptions.

The Role of Assumptions in Causal Inference

Assumptions play a dual role in causal inference: they are indispensable for causal identification and estimation, but their epistemological influence demands constant scrutiny to ensure that they do not compromise the credibility of causal estimation.

4.1.2 Assumptions in Causal Inference: Insights from Philosophy of Science

I have established so far that assumptions are essential for causal inference from the causal identification perspective. Moreover, discussions in philosophy of science also underpin the fundamental role of assumptions. In this section, some insights into assumptions from the philosophy of science will be highlighted. These insights help researchers better understand the role of assumptions in causal inference.

Implicit assumptions

Implicit assumptions are unstated premises or beliefs that underlie scientific inquiry, often operating unnoticed but exerting significant influence on theories, methods, and interpretations. Unlike explicit assumptions, which are consciously articulated and open to scrutiny, implicit assumptions remain hidden, shaping scientific practice in subtle yet profound ways. These assumptions often stem from broader paradigmatic contexts. As noted in Kuhn (1997), implicit assumptions embedded in paradigms guide what questions scientists ask and how they approach problems. For example, in consumer psychology research, many researchers implicitly assume that the findings with students or MTurkers can be generalized to the population (Pham, 2013). Similarly, in economics, the implicit assumption of universal rationality influenced decades of theoretical models before behavioral economics challenged its validity.

Because implicit assumptions are not acknowledged, they can embed biases and limit the objectivity of scientific conclusions. For example, Longino (1990) argues in *Science as Social Knowledge* the need to uncover and critique these hidden premises. By making implicit assumptions explicit and subjecting them to empirical tests or theoretical scrutiny, scientists can enhance the transparency, inclusivity, and robustness of their research.

In the context of causal inference, implicit assumptions play a crucial role but are often overlooked. For example, Difference-in-Differences (DID) analysis implicitly relies on the Stable Unit Treatment Value Assumption (SUTVA), which assumes that the treatment of one unit

does not affect the outcome of another unit. This assumption, while fundamental, is rarely acknowledged or directly examined. Yet, the violation of SUTVA can significantly bias results in settings with interference or spillover effects (e.g., education or public health policies affecting neighboring regions). Similarly, in Instrumental Variables (IV) analysis, particularly in two-stage least squares (2SLS) regression, there is an implicit assumption of linearity in the relationships between variables. If the relationship between the endogenous variable and the outcome is nonlinear, the 2SLS estimator may yield biased results, even if other IV assumptions (e.g., relevance and exclusion restriction) hold. In general, implicit assumptions can embed unnoticed biases and limit the robustness of causal conclusions. It is imperative for researchers to critically reflect on them and empirically examine them as suggested in Longino (1990).

Sufficiency and neccessity

In general, a condition is sufficient if it ensures an outcome and necessary if the outcome cannot occur without it. These ideas are key in examining causal relationships and scientific laws. For example, Hempel and Oppenheim (1948) emphasized the importance of sufficiency in their covering law model, where a scientific explanation is deemed valid if the initial conditions and laws are sufficient to deduce the phenomenon being explained. However, the distinction between sufficiency and necessity also highlights the complexity of scientific explanations: a sufficient condition might not be necessary if alternative explanations or pathways exist. The interplay between sufficiency and necessity underscores the provisional nature of scientific knowledge.

In causal inference, the distinction between sufficiency and necessity of assumptions is crucial for understanding how researchers establish causal identification. In causal identification, researchers typically aim for *sufficient assumptions*, which may not be unique or necessary. A case in point is difference-in-differences (DID) analysis, which traditionally relies on the parallel trends assumption, requiring the treated and control groups to follow the same trend in the absence of treatment. Although this assumption is sufficient to identify the causal effect, it is

often overly restrictive. Recent advances, such as methods that allow for differential trends, provide sufficient conditions for identification by assuming that the trends differ by a known and measurable amount (Roth *et al.*, 2023).

Assumptions in Causal Inference are Sufficient

In causal identification, researchers construct a set of sufficient assumptions for the causal estimand. The assumptions may not be necessary or unique. Multiple alternative sets of assumptions can all achieve identification.

One common theme in the development of causal inference is finding sufficient assumptions that are weaker or less restrictive than the previous ones. This iterative refinement of assumptions aligns with the philosophy of science's focus on simplicity and parsimony (Popper and Weiss, 1959). For example, matching methods are based on the key assumption of *conditional exogeneity* (Rubin, 1973a), which states that, conditional on observed covariates, treatment assignment is independent of potential outcomes. However, it is a strong assumption that requires that all relevant confounders be observed and considered. Recognizing the potential limitations of this assumption, researchers have sought ways to relax it by incorporating additional methods, such as doubly robust estimators (Robins *et al.*, 1994), which remains consistent if the outcome or the propensity score model, but not necessarily both, is correct.

Refutability, testability and falsifiability

The concepts of testability, falsifiability, and refutability are fundamental to understanding assumptions and the epistemology of science. These concepts relate to the criteria that assumptions (e.g., scientific hypotheses or theories) must satisfy to be considered meaningful and scientifically robust, and they have been widely discussed in philosophy of science.

Refutability is a broad concept that an assumption is refutable if it can be challenged *empirically, logically or theoretically.* Refutability

includes scenarios where an assumption might be challenged based on its internal consistency or logical coherence, even without recourse to empirical observation. **Testability** refers to the ability of an assumption to be subjected to empirical observation or experimentation. A testable assumption is one that provides observable implications and allows it to be evaluated *against empirical evidence* (Carnap, 1936). However, testability alone does not necessarily imply scientific rigor, as an assumption could be testable but trivially true or immune to meaningful challenge. Lastly, **falsifiability** is a stricter than testability. An assumption is falsifiable if it is structured in such a way that it can, in principle, be proven false by empirical evidence.

In causal inference, these concepts have profound implications due to the fundamental problem of causal inference. Since we can only observe one outcome per unit, assumptions about missing potential outcomes, such as the (conditional) unconfoundedness assumption, cannot be directly tested or falsified using observed data alone. This lack of falsifiability means that causal claims often rely on strong and untestable assumptions about the absence of unmeasured confounders.

However, while these assumptions cannot be falsified in a strict Popperian sense, they may sometimes be refuted under specific conditions when additional data become available or when theoretical insights or common knowledge contradicts the assumptions. For example, in a tiered point-based loyalty program, where consumers earn rewards by crossing predefined thresholds (e.g., 1,000 points for a discount), Regression Discontinuity Design (RDD) cannot be used to learn the effect of reaching a loyalty tier on future spending. This is because the "no perfect manipulation" can be refuted, as consumers can strategically manipulate their purchases to exceed the threshold. In general, it is important to carefully treat assumptions in empirical work using causal inference.

Key Assumptions in Causal Inference Are Untestable

In causal inference, key assumptions about potential outcomes cannot be directly tested due to the fundamental problem of causal inference.

Auxiliary assumptions and the Duhem-Quine thesis

In philosophy of science, the Duhem-Quine thesis (Duhem and Scott, 1954; Quine, 1951) highlights the interconnected nature of scientific theories and the role of auxiliary assumptions in empirical testing. This thesis argues that scientific hypotheses cannot be tested in isolation because they rely on a network of *auxiliary assumptions*, such as background theories, measurement methods, and experimental conditions. When an empirical test appears to refute a hypothesis, it is often unclear whether the hypothesis itself is flawed or if the failure lies in one of the auxiliary assumptions supporting it. In practice, the Duhem-Quine thesis underscores the complexity of scientific testing and the difficulty of pinpointing which element of a theoretical framework is responsible for inconsistency. It also highlights the importance of transparency and reflexivity in the articulation and evaluation of auxiliary assumptions, which often play a decisive role in determining the credibility and interpretation of scientific results.

The Duhem-Quine thesis has broad implications for understanding and evaluating assumptions in causal inference. It underscores the interdependence of assumptions in causal frameworks, illustrating that causal identification often relies on a network of interconnected premises. In causal inference, methods such as matching, difference-in-difference (DID), and regression discontinuity designs (RDD) depend on multiple assumptions. The Duhem-Quine thesis highlights that when a method fails to produce valid causal estimates, it can be challenging to pinpoint whether the issue lies with the primary identification assumptions (e.g., parallel trends in DID) or with auxiliary assumptions, such as measurement accuracy and model specification. Another example is the assessment of the relevance of instrumental variables. The relevance of instruments is often examined with the F-test or the Stock-Yogo weak instrument test (Stock and Yogo, 2002). If the exogeneity assumption is violated, these tests lose their meaning because the observed association between the instrument and the endogenous variable may reflect confounding rather than genuine relevance.

In general, this perspective on auxiliary assumptions encourages a holistic approach to the treatment of assumptions in causal inference.

Researchers should combine sensitivity analyses, robustness checks, and transparent reporting of assumptions to ensure credible causal conclusions. Ultimately, the Duhem-Quine thesis underscores the complexity of causal inference. The credibility of their findings depends not only on core identification assumptions, but also on the broader framework of auxiliary premises supporting their methods.

Duhem-Quine Thesis in Causal Inference

The credibility of causal conclusions depends not only on the core identification assumptions but also on the broader framework of auxiliary assumptions, and together they form a network of interdependent assumptions.

4.2 Assumptions in Stylized Research Designs

In this section, I will discuss the assumptions underlying popular causal inference methods, such as instrumental variables (IV), matching, difference-in-differences (DID), regression discontinuity design (RDD), and the synthetic control method (SCM). The underlying assumptions of these methods are critically examined through the perspectives presented in Section 4.1. In particular, I focus on three aspects in the discussion. First, for each causal inference method, I show how the assumptions are sufficient to identify the causal estimand. Second, I focus on the testability and falsifiability of assumptions and discuss whether and how common approaches to examine assumptions provide evidence for the credibility of assumptions. Third, applying the Duhem-Quine thesis, I propose that assumptions often cannot be examined in isolation as their credibility may depend on auxiliary assumptions. These ideas permeate the discussions of different methods as follows.

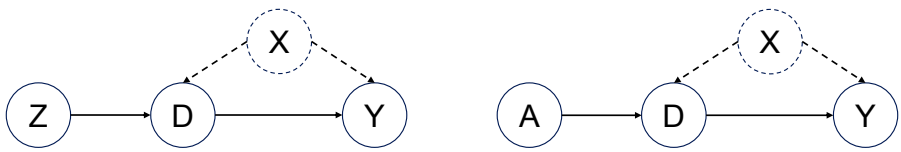
4.2.1 Instrumental Variables

Understanding IV with non-compliance

The instrumental variable (IV) strategy, as discussed in Chapter 3, is based on four key assumptions: **exogeneity**, **exclusion**, **relevance** and

monotonicity. The necessity of these assumptions is best understood via the non-compliance framework (Imbens and Angrist, 1994). In the context of experimental design, non-compliance occurs when participants do not adhere fully to their randomly assigned treatments. That is, the assignment of treatment $A_i \in \{0, 1\}$, where $A_i = 1$ means the individual is assigned to the treatment group, and $A_i = 0$ means the control group. However, due to non-compliance, the actual receipt of treatment $D_i \in \{0, 1\}$ may not align perfectly with the assignment, with $P(A_i = D_i) \in (0, 1)$. In comparison, in a standard complete randomized experiment, the assignment is always equal to the treatment received $P(A_i = D_i) = 1$. The non-compliance problem bears the same structure as the general instrumental variable strategy, as illustrated in Figure 4.1. Therefore, the insights from studying the non-compliance problem applies to the general instrumental variable strategy.

Figure 4.1: DAGs of IV Strategy and Non-compliance



(a) General Instrumental Variable Strategy **(b)** Non-compliance in Experiments

To understand why IV assumptions are needed, we start with the insightful categorization of compliance types of people in experiments (Imbens and Angrist, 1994). In general, people can be classified into four **latent types**[1]:

$D(A_i = 0)$	$D(A_i = 1)$	Compliance Types
0	0	Never-taker (nt)
0	1	Complier (co)
1	0	Defier (de)
1	1	Always-taker (at)

The potential outcomes for the final variable of interest, $Y_i(A_i, D_i)$,

[1]Note that the type of any particular individual cannot be observed by researchers. The types refer to the finite population, such that the types of people would differ in two experiments of different finite populations.

depend both on the assigned treatment A_i and whether the individual complies with the assignment, represented by D_i. This results in four possible potential outcomes based on combinations of A_i and D_i as shown in the table below, with $D_i(0) = D(A_i = 0)$ and $D_i(1) = D(A_i = 1)$. Since only one of these potential outcomes is observed for each individual, depending on their realized A_i and D_i, the other three outcomes are missing.

Assignment	Treatment	Potential Outcome
$A_i = 0$	$D_i(0) = 0$	$Y_i(0, 0)$
$A_i = 0$	$D_i(0) = 1$	$Y_i(0, 1)$
$A_i = 1$	$D_i(1) = 0$	$Y_i(1, 0)$
$A_i = 1$	$D_i(1) = 1$	$Y_i(1, 1)$

A step-by-step approach to IV assumptions

Next, I will follow the logic from the **law of decreasing credibility** (Manski, 2003) and gradually add the assumptions for IVs. This process helps to illuminate the roles of different assumptions. First, assume that the only assumption we have is that the assignment A_i is exogenous, which is generally true if A_i is randomized. With the **exogeneity** assumption, we can directly identify the effect of A_i on D_i and Y_i, as seen from the DAG in Figure 4.1. These are known as the intent-to-treat effect (ITT), noted as $\text{ITT}_D = P(D \mid A)$ and $\text{ITT}_Y = P(Y \mid A)$. These two quantities can be directly obtained from the data $\Phi = P(A, D, Y)$. The interpretation of the ITT is the effect of assignment A on the outcomes.

To better understand this, I will use loyalty programs as an example. The effects of loyalty programs on consumer behaviors, such as purchases and brand loyalty, have been studied extensively in marketing (e.g., Bolton *et al.*, 2000; Sharp and Sharp, 1997). More recently, marketing researchers have begun to use field experiments to investigate the effects of loyalty programs (e.g., Wang *et al.*, 2016). Suppose that an airline company intends to study the effect of their loyalty program on ticket revenue from customers. The design of the experiment is as such: the company sends randomly selected customers an incentive package (i.e., small gifts) for joining the loyalty program. The company

realizes the non-compliance issue in the experiment and estimates the ITT on revenue. The interpretation of ITT on revenue is the effect of the marketing campaign ("sending small gifts") on revenue, but not the loyalty program. So, the question remains, can we learn anything about the effect of loyalty program? To do so, we need to make more assumptions.

To proceed to the next step, notice that the overall ITT_Y can be decomposed into:

$$\begin{aligned}\text{ITT}_Y = &\text{ITT}_Y^{\text{co}} \times \pi_{\text{co}} + \text{ITT}_Y^{\text{nt}} \times \pi_{\text{nt}} \\ &+ \text{ITT}_Y^{\text{at}} \times \pi_{\text{at}} + \text{ITT}_Y^{\text{de}} \times \pi_{\text{de}},\end{aligned} \tag{4.1}$$

where $\text{ITT}_Y^{\text{g}} = E\left(Y_i(1, D_1^i) - Y_i(0, D_0^i) \mid i \in g\right)$ and $\pi_g = P(i \in g)$. Next, the **exclusion** assumption is added on top of the **exogeneity** assumption. The exclusion assumption implies that the potential outcomes of Y is the same given the same values of D, such that $Y_i(0,0) = Y_i(1,0)$ and $Y_i(0,1) = Y_i(1,1)$. Under the **exclusion** assumption, for the always-takers and never-takers, because their actual receipts of the treatment are always the same, irrespective of the assignment A_i, the following equations hold:

$$\begin{cases}\text{ITT}_Y^{\text{at}} &= E\left(Y_i(1,1) - Y_i(0,1) \mid i \in at\right) = 0 \\ \text{ITT}_Y^{\text{nt}} &= E\left(Y_i(1,0) - Y_i(0,0) \mid i \in nt\right) = 0\end{cases} \tag{4.2}$$

This simplifies Equation 4.1 into the following,

$$\text{ITT}_Y = \text{ITT}_Y^{\text{co}} \times \pi_{\text{co}} + \text{ITT}_Y^{\text{de}} \times \pi_{\text{de}} \tag{4.3}$$

However, this equation does not communicate much since it has four unknowns $\{\text{ITT}_Y^{\text{co}}, \text{ITT}_Y^{\text{de}}, \pi_{\text{co}}, \pi_{\text{de}}\}$. Following the same logic as above, extra assumptions need to be included. Next, let us add the **monotonicity** assumption that the value of D increases or decreases with A. The **monotonicity** assumption (a.k.a. no-defiers) rules out the existence of defiers, which implies the value $\pi_{\text{de}} = 0$. The decomposition in Equation 4.3 is further simplified into,

$$\text{ITT}_Y = \text{ITT}_Y^{\text{co}} \times \pi_{\text{co}} \tag{4.4}$$

It also turns out under the **monotonicity** assumption, the proportion of compliers π_{co} is equal to the intent-to-treat on D or ITT_D (see Imbens

and Angrist, 1994, for more details). Finally, the **relevance** assumption ensures that intent-to-treat on the receipt of the treatment $\text{ITT}_D \neq 0$. Therefore, given the four assumptions: **exogeneity**, **exclusion**, **relevance** and **monotonicity**, the complier (local) average treatment effect can be learned from the data $\Phi = P(A, D, Y)$.

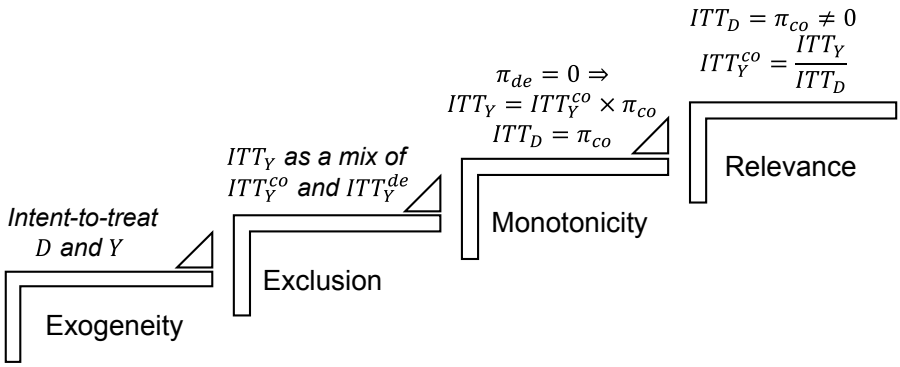
$$\text{ITT}_Y^{\text{co}} = \frac{\text{ITT}_Y}{\text{ITT}_D} \tag{4.5}$$

Summary

In summary, insights into these four assumptions emerge from the above identification process. First, there is a hierarchy of the four assumptions by following the logic of the law of decreasing credibility, as illustrated in Figure 4.2. The starting point is the **exogeneity** assumption, enabling the identification of the intent-to-treat effect. Without the **exogeneity** assumption, one learns *nothing* about treatment effects from the data $\Phi = P(A, D, P)$ alone. The addition of **exclusion**, equates ITT_Y to a weighted average of ITT_Y^{co} and ITT_Y^{de}, although it does not lead to new identification results. The **monotonicity** assumption further rules out the existence of defiers and equates ITT_Y to the weighted value of ITT_Y^{co} and allows the identification of π_{co} via ITT_D. Finally, the **relevance** assumption gives the mathematical guarantee to solve Equation 4.4.

Second, the first two assumptions **exogeneity** and **exclusion** are about potential outcomes, which makes them untestable and therefore unfalsifiable (Proposition 4.3. The other two assumptions **monotonicity** and **relevance** are NOT about the potential outcomes. **Monotonicity** is to rule out the existence of defiers, but it is still conceptual by nature and untestable as researchers do not observe the types of people. Therefore, researchers must justify this assumption with their knowledge of the empirical settings. The discussion often revolves around how the assignment does not lead people to "rebel." The **relevance** assumption, on the other hand, is statistical and can be tested *only if* the first two assumptions hold, following the logic of Proposition 4.4.

Figure 4.2: The Hierarchy of IV Assumptions



4.2.2 Matching

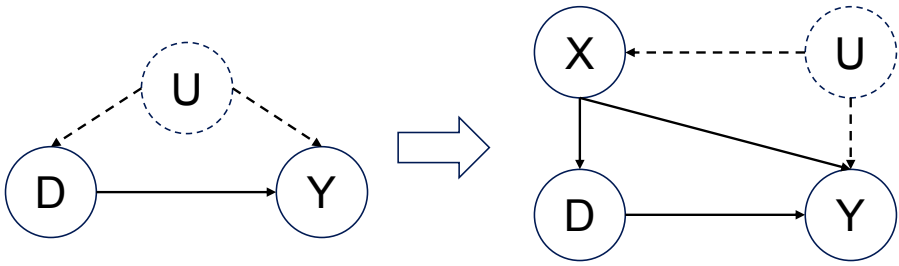
This technique has been available since it was introduced by Rosenbaum and Rubin (1983a). In marketing, there are many applications. For example, Huang *et al.* (2012) investigated the effects of Walmart's growth on suppliers' profits; Sudhir and Talukdar (2015) used it to address a paradox about the adoption of IT by Indian retailers, using data from a comprehensive survey; De Haan *et al.* (2018) examined the influence of device switching on online purchasing behavior; and Eggert *et al.* (2019) investigated how gift purchases affect customer perceptions and future purchasing habits.

Matching as an artificial block design

The fundamental assumption in matching designs is the **conditional exogeneity**, which states that the treatment D is unconfounded conditional on the observed set of variables X (Rubin, 1973a), as shown the DAG in Figure 4.3. I view matching as a specific research design based on the general **conditioning strategy** discussed in Chapter 3. In addition, in the following discussion, I focus on the base version of matching instead of the propensity score matching, but all the conclusions apply as the propensity score is a sufficient statistics of X (Rosenbaum and Rubin, 1983a).

From an experimental design perspective, matching draws similari-

Figure 4.3: The DAG for Matching



ties with block randomization. I will explain the intuitions of matching using the ideas in block randomization. Block randomization aims to control for variability among participants by grouping them into a finite number of homogeneous blocks based on certain characteristics (e.g., genders), denoted as X (the blocking variables). Within each block, treatment $D \in \{0, 1\}$ is randomly assigned. Therefore, block randomization can be viewed as conducting a separate randomized experiment within each block. For each block b, the average treatment effect (ATE) can be estimated as the difference in expected outcomes Y between the treated and control groups, conditional on the blocking variable X_b. Mathematically, the ATE for block b is:

$$\text{ATE}_b = E(Y(1) \mid D = 1, X = X_b) - E(Y(0) \mid D = 0, X = X_b). \quad (4.6)$$

Since treatment is randomized within each block, the causal effect of D on Y can be identified independently for each block. To obtain the overall ATE for all blocks, the ATEs of individual blocks are combined as a weighted average. Let p_b represent the proportion of individuals in block b, then the overall ATE is given by $\text{ATE} = \sum_b p_b \text{ATE}_b$.

However, in a matching design, researchers do not control and observe these "blocks." On the other hand, under the **conditional exogeneity** assumption, the set of conditioning variables X are observed and render the treatment D conditional exogenous. So, researchers can "reverse engineer" the blocks and obtain the ATE in a way similar to block randomization, by first obtaining the ATEs of "blocks" and then combing them as shown in Equation 4.6. In fact, the subclassification method in matching works exactly as this (see p.382 of Imbens and

Rubin, 2015).

$$\begin{cases} \text{ATE}_x & = \text{E}(Y \mid D = 1, X = x) - \text{E}(Y \mid D = 0, X = x) \\ \text{ATE} & = \sum_x p_x \text{ATE}_x \end{cases} \tag{4.7}$$

The **conditional exogeneity** assumption ensures the following equalities, similar to a randomized experiment at each realization of $X = x$.

$$\begin{cases} \text{E}(Y(1) \mid X = x) & = \text{E}(Y(1) \mid D = 1, X = x) \\ \text{E}(Y(0) \mid X = x) & = \text{E}(Y(0) \mid D = 1, X = x) \end{cases} \tag{4.8}$$

Aside from the **conditional exogeneity** assumption, matching also requires an assumption called the **common support assumption**. Formally, the assumption states that the probability of treatment conditional on X is strictly bounded between 0 and 1 or $P(D = 1 \mid X) \in (0, 1)$. The implication of this assumption is that the treatment and control group are overlapped in terms of X. That is, researchers can always find people bear similar characteristics in the treatment and control group. Mathematically, this is a "regularity condition," which ensures that conditional expectations $\text{E}(Y(1) \mid D = 1, X = x)$ and $\text{E}(Y(0) \mid D = 0, X = x)$ can be learned from data. In contrast, if $P(D = 1 \mid X = x') = 1$, then $\text{E}(Y(0) \mid D = 0, X = x')$ is not identified as $P(D = 0 \mid X = x') = 0$.

Concluding remarks on assumptions in matching

In sum, the two assumptions of the matching designs ensure the identification of ATE from the data $\Phi = P(D, Y, X)$. The assumption of **conditional exogeneity** is about potential outcomes and therefore untestable and unfalsifiable. The **common support** assumption, on the other hand, a regularity condition that ensures that the conditional expectation can be learned from the data. Intuitively, it ensures that treated and control units have comparable counterparts across the entire range of covariates. The **common support** assumption can be post-hoc guaranteed by trimming samples from the treated and control units. However, a reminder is that trimming may also lead to shifts in finite

populations [2].

Lastly, I will discuss the role of the widely adopted **balance tests** in matching designs. As shown in Abadie and Imbens (2006), discrepancies in X between the treated and control groups will generally induce biased causal estimates. In practice, research using matching often seeks to achieve a good balance of covariates X between the treated and control groups. It is important to note that balance tests or other tests of equivalence of the treatment and control groups do not provide convincing evidence of the credibility of the two matching assumptions. In fact, efforts to balance covariates only reduce the bias if conditional exogeneity and the common support assumption hold. In practice, balance tests should be treated as an assessment of the quality of matching efforts. From the identification-estimation flow (see Figure 3.1), the bias from imbalanced covariates or incomplete matching (Rosenbaum and Rubin, 1985) is statistical in nature, reflecting the limitations of sample sizes and model specifications[3].

4.2.3 Difference-in-Difference

The difference-in-difference (DID) design is a type of event study design that exploits time variation before and after the start of an intervention (Miller, 2023). DID has been widely adopted in marketing research to study various problems, such as the adoption of a mobile shopping app by retail customers (Narang and Shankar, 2019), the introduction of payment disclosure laws in prescriptions by physicians (Guo *et al.*, 2021) and the effect of augmented reality in sales of online retailers (Tan *et al.*, 2022).

To better understand the advantages of DID, I will first discuss the traditional event study used frequently in finance (MacKinlay, 1997), strategy (Park, 2004), and marketing (Sorescu *et al.*, 2017). In a

[2]A finite population refers to a fixed, well-defined set of individuals or units from which we want to draw causal conclusions. The key feature of a finite population is that it is not an infinitely large or hypothetical group; instead, it consists of a specific number of individuals or units that are of interest for the analysis.

[3]The argument applies to propensity scores as the true functional form of the propensity score is unknown. Parametric specification and the finite sample can lead to bias in estimated propensity scores

traditional event study, a treatment or intervention occurs at a point in time in a long time series. The long time series allow researchers to observe the evolution of treatment effects before and after treatment. Suppose that there are N units observed over T time periods, where a treatment D occurs at a specific time t_1 for *all* N *units*. The goal is to construct credible counterfactual outcomes $Y_{t>t_1}(0)$ for the N units after the treatment using pre-treatment time series $Y_{t\leq t_1}(0)$. With the constructed counterfactual outcomes, the treatment effect is defined as $\mathrm{E}(Y_{t>t_1}(1) - \hat{Y}_{t>t_1}(0))$. However, the traditional event study requires a strong assumption for identification: the time series of the outcome Y would be **stationary** had the treatment not occurred, which is difficult to hold in reality.

The basic setup of DID

Unlike the traditional event study, a DID design allows the existence of a control group who are never treated. This enables researchers to relax the strong assumption of stationarity in a traditional event study. In a canonical DID setup, suppose that there are N units observed across $T = 2$ time periods. Let D_{it} represent the binary treatment indicator, where $D_{it} = 1$ for treated units and $D_{it} = 0$ for control units. The (potential) outcomes are represented by Y_{it} for unit i at time t. Typically, researchers are interested in estimating the Average Treatment Effect on the Treated (ATT), defined as

$$\tau_{ATT} = E[Y_{i2}(1) - Y_{i2}(0) \mid D_{i2} = 1],$$

that captures the treatment effect at time $t = 2$ for the treated group.

A point to emphasize here is the typical assumptions (parallel trend and no anticipation) in DID enable researchers to identify the ATT. Therefore, if ATE is needed, the typical assumptions are insufficient. Identifying ATT is reasonable in many economic studies, because policymakers and economists are primarily interested in understanding the welfare of individuals, firms or regions directly affected by a policy or intervention, rather than hypothetical effects on the entire population (Lechner *et al.*, 2011). Therefore, the estimand from DID needs to be interpreted with caution, especially given the treatment group usually

self-select themselves into treatments. The learned effect thus only applies to the finite population of the treatment group. The credibility of any argument that intends to generalize to the super-population should be up to scrutiny.

The assumptions for identifying ATT

To understand the assumptions in DID, let us again follow the **law of decreasing credibility** and start with no assumptions. Without further assumptions, the data Φ from the above data generation process contains the following observed outcomes for the treatment and control group. For the treatment group, the potential outcome of no treatment is observed in the before period and that of treatment in the after period, with $\Phi_{\text{treatment}} = \{Y_{i1}(0), Y_{i2}(1)\}$. In comparison, for the control group, the potential outcome of no treatment is observed in both the before and after period, since the control group is never treated, with $\Phi_{\text{control}} = \{Y_{i1}(0), Y_{i2}(0)\}$.

Table 4.1: Potential Outcomes for the Treated and Control Group in DID

Treated	$T=1$	$T=2$
$D=0$	$Y_{i1}(0)$	~~$Y_{i2}(0)$~~
$D=1$	~~$Y_{i1}(1)$~~	$Y_{i2}(1)$

(a) Treated Group

Control	$T=1$	$T=2$
$D=0$	$Y_{i1}(0)$	$Y_{i2}(0)$
$D=1$	~~$Y_{i1}(1)$~~	~~$Y_{i2}(1)$~~

(b) Control Group

The causal estimand for DID is the ATT, which is defined as $\tau_{\text{ATT}} = \text{E}(Y_{i2}(1) - Y_{i2}(0) \mid D_{i2} = 1)$. In comparison, ATE is defined as $\tau_{\text{ATE}} = \text{E}(Y_{i2}(1) - Y_{i2}(0))$. ATE identification, as discussed in Chapter 2 requires the exogeneity assumption that treatment at $t = 2$ is unconfounded. As a matter of fact, if the exogeneity assumption holds, it is unnecessary to observe data in the pre-treatment period. If we take a closer look at ATT, the first part $\text{E}(Y_{i2}(1) \mid D_{i2} = 1)$ is the result for the treatment group at $t = 2$. The second part $\text{E}(Y_{i2}(0) \mid D_{i2} = 1)$ is unobserved. The key question therefore becomes to find sufficient assumptions to learn $\text{E}(Y_{i2}(0) \mid D_{i2} = 1)$ from the data. It is obvious that the exogeneity assumption makes $\text{E}(Y_{i2}(0) \mid D_{i2} = 1) = \text{E}(Y_{i2}(0) \mid D_{i2} = 0)$, and

therefore identifies the ATT. However, given the observation of pre-treatment data, this assumption can be relaxed.

It turns out the observation of pre-treatment data allow us to have a weaker version of exogeneity by assuming **the parallel trend** as below,

$$\mathrm{E}(Y_{i2}(0) - Y_{i1}(0) \mid D_{i2} = 1) = \mathrm{E}(Y_{i2}(0) - Y_{i1}(0) \mid D_{i2} = 0) \tag{4.9}$$

The parallel trend assumption essentially says that the before-after differences in potential outcomes of the treatment and control group are equal. Or, in other words, **the before-after difference in potential outcomes is unconfounded**. The assumption allows the identification of the unknown potential outcome $\mathrm{E}(Y_{i2}(0) \mid D_{i2} = 1)$ from data as shown below,

$$\begin{aligned}\mathrm{E}(Y_{i2}(0) \mid D_{i2} = 1) = & \underbrace{[\mathrm{E}(Y_{i2}(0) \mid D_{i2} = 0) - \mathrm{E}(Y_{i1}(0) \mid D_{i2} = 0)]}_{\text{Observed for Control in } t=1,2} \\ + & \underbrace{\mathrm{E}(Y_{i1}(0) \mid D_{i2} = 1)}_{\text{Observed for Treated in } t=1}\end{aligned} \tag{4.10}$$

Another way to understand the **parallel trend** assumption is from the selection bias perspective as shown in Equation 4.11, a rearrangement of terms in 4.9. The parallel trend assumption is equivalent to assuming that the selection bias of potential outcomes under no treatment is constant in the before-and-after period.

$$\begin{aligned}&\underbrace{\mathrm{E}(Y_{i2}(0) \mid D_{i2} = 1) - \mathrm{E}(Y_{i2}(0) \mid D_{i2} = 0)}_{\text{Selection bias in } t=2} = \\ &\underbrace{\mathrm{E}(Y_{i1}(0) \mid D_{i2} = 1) - \mathrm{E}(Y_{i1}(0) \mid D_{i2} = 0)}_{\text{Selection bias in } t=1}\end{aligned} \tag{4.11}$$

Finally, aside from the **parallel trend** trend assumption, the identification of ATT requires an important but often implicit assumption. This assumption states that the treatment has no causal effect prior to its actual implementation. Formally, the **no anticipation assumption** is expressed as below,

$$Y_{i1}(0) = Y_{i1}(1), \forall i \in \{i \mid D_{i2} = 1\}. \tag{4.12}$$

This is essential for identification, as the changes in the outcome for the treated group between periods 1 and 2 could be confounded by

the anticipatory effect (Malani and Reif, 2015). The **no anticipation** assumption generally receives less attention in empirical research, despite the ample evidence that consumers could be forward-looking in their decisions (e.g., Ching and Osborne, 2020; Villas-Boas, 2004).

The strength and weakness of DID assumptions

The **parallel trend** assumption is weaker than the exogeneity assumption. From Equation 4.11, under the assumption of parallel trend, the selection bias is allowed but is assumed to be constant over time. The existence of the pre-treatment period allows researchers to control for the selection bias and partial it out from the final treatment effect. This is an important point of view on DID design. In fact, the view of selection bias inspires the sensitivity analysis on the assumption of parallel trend in Rambachan and Roth (2023). The **no anticipation** assumption is essentially an exclusion condition to rule out the possible confounding effect from anticipating the treatment. In many contexts, the credibility of **no anticipation** is up to debate. Therefore, empirical research should discuss its credibility using institutional knowledge. In addition, if conditions allow, a test of anticipatory behavior based on the announcement of the implementation of the treatment (Gruber and Köszegi, 2001). Finally, both the **parallel trend** and **no anticipation** assumption are about potential outcomes and therefore untestable.

On the surface, compared with the settings using only cross-sectional comparison (e.g., regression adjustment or matching), the DID design imposes weaker assumptions due to its focus on the ATT instead of ATE and the use of observations across time. However, exploiting variations in time is both "a blessing and a curse". This is best understood with a two-way fixed effect (TWFE) model, a workhorse in DID analysis, especially for multiple time periods (Lechner *et al.*, 2011).

$$y_{it} = \alpha_i + \lambda_t + \tau D_i \times T_t + \varepsilon_{it} \tag{4.13}$$

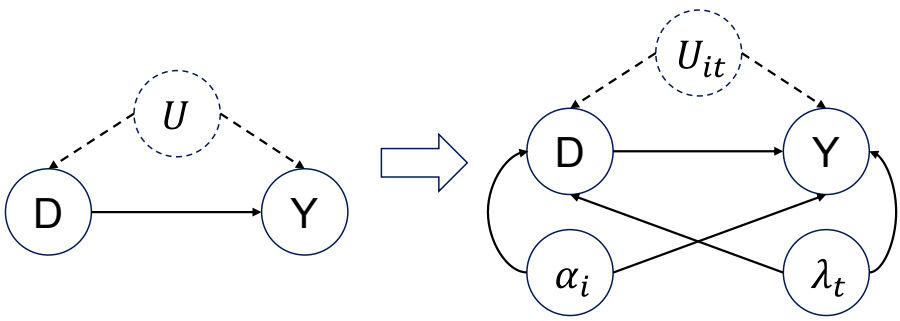
In Equation 4.13, α_i and λ_t are the individual- and time-fixed-effects. D_i is the treatment dummy and T_t the before-after dummy. The estimation of τ under the standard assumptions of DID is the ATT. The **parallel trend** assumption for the TWFE model renders the following

relationship for the error term,

$$\mathrm{E}(\varepsilon_{i2} - \varepsilon_{i1} \mid D_{i2} = 1) = \mathrm{E}(\varepsilon_{i2} - \varepsilon_{i1} \mid D_{i2} = 0) \tag{4.14}$$

That is, the before-after differences in the unobservable ε of the treatment and control groups are the same. This also implies that the DID design is subject to bias from time- and individual-varying unobservables. The DAG of DID in Figure 4.4 illustrates this point. In conclusion, although the DID design manages to relax the exogeneity assumption with the focus on ATT and the use of time series, it also opens to the bias that can be induced by time, especially the individual and time-varying unobservables.

Figure 4.4: The DAG for DID



4.2.4 Regression Discontinuity Design

The Regression Discontinuity Design (RDD) is a quasi-experimental method used to estimate causal effects when there is a clear cut-off or threshold determining eligibility for treatment. This design exploits situations where a continuous variable (that is, the variable **running**) determines treatment assignment, with individuals just above the threshold receiving treatment and those just below it not receiving it. RDD is popular in empirical research in part because such discontinuities frequently occur in the real world due to policies, regulations, or eligibility criteria. For examples, in marketing research, Narayanan and Kalyanam (2015) relies on the discontinuities in the rankings in search ads auctions to identify position effects in search advertising. And Zhong (2022) uses

the discontinuities in the scoring of online sellers to study the effect of assigning symbols to sales on retailers' strategic responses.

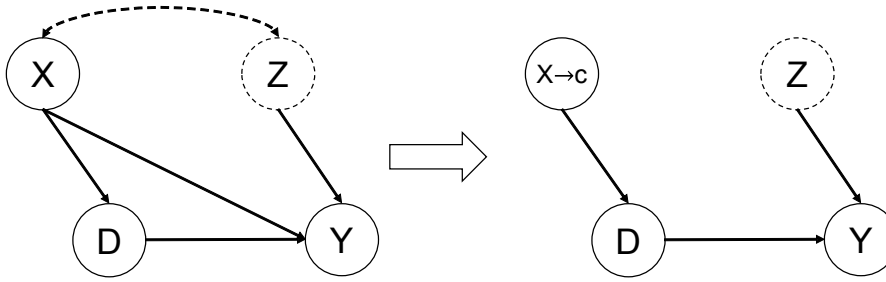
What can we learn from RDD?

For a sharp regression discontinuity design (RDD), suppose that we have a treatment $D \in \{0, 1\}$ that is determined by the (running or forcing) variable X. We have $D = 1$, if $X \geq c$ and $D = 0$, if $X < c$, where c is the predetermined cutoff point. We want to know how treatment D influences an outcome Y. Therefore, two questions naturally follow: *What can we learn about the treatment effect of D*? And *what assumptions need to be maintained*? To answer the first question, I will first discuss RDD using the ideas from matching.

Given the data-generating process of RDD, treatment D depends only on the running variable X. In the classic matching approach, we look for units in the treatment and control groups with the same propensity to treat with $p(X_i|D_i = 1) = p(X_j|D_j = 0)$. For RDD, the propensity score degrades to a stepping function of 0-1 and the only point where two people in the treatment vs. control are "equivalent" is the cutoff $X = c$. Given that X is continuously distributed, the probability of $X = c$ shrinks to zero with $P(X = c) = 0$. Therefore, RD design can be seen as a limiting case of matching at the cutoff $X = c$. The usual matching methods, therefore, fail to apply because the common support assumption is not satisfied. In a limiting sense, we can only identify the effect of treatment exactly at the cutoff point. This is essentially what we can learn from the RDD - **the treatment effect at the cutoff**. The DAG of RDD in Figure 4.5 illustrates the intuitions.

The answer to the first question informs us about the causal estimand of RDD - the treatment effect at the cutoff. The second question still remains open: *What assumptions are needed to learn the treatment effect at the cutoff*? In RDD literature, the most frequently used term for the assumptions are "continuity" and "no perfect manipulation" (Cattaneo and Titiunik, 2022; Lee and Lemieux, 2010). In the following sections, I will discuss the two assumptions from the potential outcome and local experiments perspective, and show that these two assumptions

Figure 4.5: The DAG for RDD



are mathematically equivalent.

The potential outcome perspective

As discussed above, the matching method does not apply to RDD. It is tempting to run an estimating equation like the following,

$$Y = \alpha + \tau D + \beta X + \varepsilon \tag{4.15}$$

However, this equation cannot give an unbiased estimation of the treatment effect τ. This is because although the treatment D is determined by the running variable X, it is possible that other variables can be related to X (and hence D) and Y. For example, if X is the grades and Y is the job earnings, the innate abilities of people would also be related to D the cum laude and the outcome Y the earnings. So, the question is *when D becomes (almost) exogenous*?

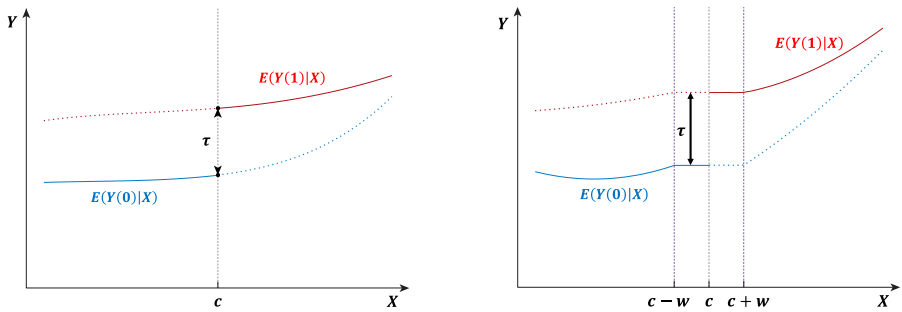
To answer this question, imagine that we focus on a narrow neighborhood around the cutoff c with $(c - \Delta c, c + \Delta c)$. Using a narrower neighborhood would produce more credible $\hat{\tau}$, as two sets of people in the narrower neighborhood above and below the cutoff are more similar to each other. If we formalize the intuition using potential outcomes, we define the treatment effect at the cutoff as:

$$\tau(c) = \lim_{\Delta c \downarrow 0} E(Y_i | X_i = c + \Delta c) - \lim_{\Delta c \uparrow 0} E(Y_i | X_i = c + \Delta c) \tag{4.16}$$

For $\tau(c)$ to exist, a sufficient condition is that the limit of the potential outcomes exist. A necessary and sufficient condition for potential outcomes to exist in c is that both potential outcomes are continuous

in c. Usually, we adopt a sufficient assumption that **the potential outcomes are continuous everywhere**. This is generally known as the **continuity assumption**. Figure 4.6[4] illustrates the potential outcome perspective on RDD. The central idea is to "extrapolate" the potential outcome functions $\mathrm{E}\left(Y(1) \mid X\right)$ and $\mathrm{E}\left(Y(0) \mid X\right)$ to the cutoff point c.

Figure 4.6: The Potential Outcome vs. Local Experiment Perspective on RDD



(a) Potential Outcome

(b) Local Experiment

The local experiment perspective

In a randomized experiment, people are typically divided into treatment and control groups based on a randomly generated number, ν. For example, ν can be a standard uniform random variate and subjects are assigned to the treatment group if $\nu > 0.5$. Therefore, a randomized experiment can be seen as a RD design where $X = \nu$ is the running variable and $c = 0.5$ the cutoff. The **only** difference is that the running variable is uncorrelated with the potential outcomes. Therefore, in a randomized experiment, the expected potential outcomes $\mathrm{E}(Y_i^1|X)$ and $\mathrm{E}(Y_i^0|X)$ become constants with respect to X. Randomization naturally leads to the continuity of potential outcomes. However, in a RD design, the running variable is usually correlated with the potential outcomes. Therefore, a simple comparison $\mathrm{E}(Y_i^1|X)$ and $\mathrm{E}(Y_i^0|X)$ no longer yields a consistent estimate of the treatment effect.

[4]The figure is reproduced from Figure 1 of Cattaneo and Titiunik (2022).

To formalize the idea, let us assume, without loss of generality, the data-generating process for the RD is captured in the following linear equations (Lee and Lemieux, 2010):

$$\begin{aligned} Y &= \tau D + \delta_1 Z + U \\ D &= 1(X \geq c) \\ X &= \delta_2 Z + V \end{aligned} \tag{4.17}$$

In Equation 4.17, Z represents the observed characteristics and U and V the unobserved ones. Note that Z can be endogenously determined as long as the determination is made prior to treatment. The system of equations resembles a regression with an endogenous dummy variable D, except that we know how D is determined. This allows us to relax strong assumptions. For example, variables $\{Z, U, V\}$ can be arbitrarily correlated and no exclusion restriction is needed for identification, different from the sample selection model (Heckman and Honore, 1990).

From the system of equations, the distribution of X (or V) fully depends on $\{Z, U\}$ and the individual heterogeneity is fully described by $\{Z, U\}$. That is, any two people with the same $\{Z, U\}$ will have the same X. If an individual has *complete and exact control* over X (and thus $\{Z = z, U = u\}$), we would observe no people below the cutoff point, assuming the treatment is beneficial for people. In comparison, if people *cannot precisely control* X and therefore there are stochastic errors in the assignment variables, we would expect the density of X (and V), conditional on Z and U to be continuous at the cutoff c. Therefore, continuity is a consequence of people having imprecise control over X or "local randomization." A formal proof is as below:

$$f(Z = z, U = u \mid X = x) = f(x \mid Z = z, U = u) \frac{P(Z = z, U = u)}{f(x)} \tag{4.18}$$

The continuity of the right-hand side depends on the continuity of $f(x \mid Z = z, U = u)$. And the function of the outcome Y is continuous *if and only if* $f(Z = z, U = u \mid X = x)$ is continuous. Under **no perfect manipulation**, $f(x \mid Z = z, U = u)$ must be continuous, and therefore $f(Z = z, U = u \mid X = x)$ and Y. The reverse is also true. Therefore, if individuals have cannot perfectly manipulate X, the

treatment is as good as randomly assigned around the cutoff and the function of the outcome Y is continuous. On the flip side, the continuity of Y also implies non-perfect manipulation. In conclusion, the **local randomization** and the **continuity** assumption are equivalent.

Given the local randomization condition, we will focus on a small neighborhood of the cutoff c to obtain samples for the local experiment.

$$\begin{aligned}\hat{\tau}(c) &= E(Y \mid X \in (c, c+\Delta c)) - E(Y \mid X \in (c-\Delta c, c)) \\ &= \tau(c) + \int_{z,u} (\delta_1 z + u) dF(z, u \mid X \in (c, c+\Delta c))) \\ &\quad - \int_{z,u} (\delta_1 z + u) dF(z, u \mid X \in (c-\Delta c, c))\end{aligned} \tag{4.19}$$

The bias disappears when the last terms cancel out. In the limiting sense, as $f(z, u \mid X)$ is continuous, if $\Delta c \to 0$, then the bias becomes zero. Empirically, we need to focus on a narrow neighborhood of c for the bias to asymptotically become zero.

A synthesis of the two perspectives

The two perspectives have different implications for how to assess the validity of RD designs. The potential outcomes perspective focuses on the continuity of potential outcomes. Since potential outcomes are missing due to the fundamental problem of causal inference, researchers often set to test continuity of variables that are related to the outcome but not influenced by the treatment. On the other hand, the local randomization condition implies that covariates are balanced (or continuous) across the cutoff. Therefore, balance tests for covariates Z can be used to examine the validity of RDDs, similar to a complete randomization. Another way of validating the local randomization condition (or no perfect manipulation) is to look at the distribution of X directly around the cutoff, which is known as the McCrary density test (McCrary, 2008).

Finally, it is also important to understand that adding pre-determined covariates to RD analysis is unnecessary or irrelevant for the identification of the estimated treatment effect, as the local randomization implies the balance of covariates. However, in practice, adding some pre-treatment covariates may improve the statistical inference and in-

crease one's statistical power, especially if the covariates are correlated with the potential outcomes (Imbens and Lemieux, 2008).

4.2.5 Synthetic Control Method

The synthetic control method (SCM), as discussed in Abadie *et al.* (2010) and Abadie (2021), is an econometric tool designed to evaluate the effects of policy interventions, particularly when those interventions impact aggregate entities such as cities, regions, or countries. This methodology constructs a synthetic control *unit* - a weighted average of unexposed units ("the donor pool") - that approximates the characteristics and pre-intervention outcomes of the treated unit. The method is designed as an improvement over traditional comparative case studies, which often rely on informal selection of comparison units. Synthetic controls address this by leveraging a formalized, data-driven process to construct a comparison unit that more closely represents the treated unit.

At its core, the SCM aims to learn how the outcome of interest (e.g., GDP, unemployment rates) for the treated unit would have evolved in the absence of the intervention. In other words, the SCM focuses on one treated unit and infers *the individual treatment effect* (ITE) of the unit with information from multiple control units. The objective of learning ITE fundamentally differentiates SCM and DID [5]. Due to *the fundamental problem of causal inference*, researchers should approach the results of SCM with caution. ITE is not identified even with the treatment randomized to units (Bottmer *et al.*, 2024).

The basics of SCM

Suppose that we observe $J+1$ units with the unit $i=1$ treated and the remaining J units untreated. The J untreated units are often called the "donor pool". In addition, both pre- and post-treatment periods are observed, with T_0 pre-treatment periods out of total T periods. Let Y_{it} represent the observed outcome for unit i at time t, and let Y_{it}^0 and Y_{it}^1 denote the potential outcomes without intervention and intervention,

[5] In recent years, researchers try to apply the SCM idea in DID method, for example, Arkhangelsky *et al.* (2021).

respectively. The treatment effect for the treated unit (unit 1) at time $t > T_0$ is given by $\tau_{1t} = Y_{1t}^1 - Y_{1t}^0$.

Since Y_{1t}^0 is unobservable, the potential outcome is estimated as $\hat{Y}_{1t}^0 = \sum_{j=2}^{J+1} w_j Y_{jt}$, where w_j are nonnegative weights that sum to one. These weights are chosen to minimize the discrepancy between the treated unit and the synthetic control during the pre-treatment period, solving the optimization problem

$$\min_W \|X_1 - X_0 W\|^2, \text{ subject to: } \sum_{j=2}^{J+1} w_j = 1 \text{ and } w_j \geq 0, \quad (4.20)$$

where X_1 is the vector of pre-treatment predictors for the treated unit, X_0 is the matrix of predictors for the donor pool, and $W = (w_2, w_3, \ldots, w_{J+1})^\top$ is the vector of weights. The predictors are for the untreated outcomes in the post-treatment period. After constructing the synthetic control, the effect of treatment is estimated at $\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^N$ for all $t > T_0$.

The assumptions of SCM

From the basic setup of SCM, I have established that SCM aims to identify the individual treatment effect by using the control units to construct the counterfactual for the treated unit. Although SCM has similarities to DID, it fundamentally differs from DID in its causal estimand. Therefore, the identification assumptions also differ from DID. As a starting point, SCM requires much stronger assumptions compared with DID for nonparametric identification. Using the formalization of identification in 3, the focal estimand $\theta = \tau_{1t}$ and the data are $\Phi = \{Y, X_0, X_1\}$, where Y is the data matrix of the outcomes for all units and time periods, and X the pre-treatment predictors.

Abadie *et al.* (2010) assumes the potential outcomes are generated by a linear factor model with interactive fixed effects (Bai, 2009) and discusses the identification based on the model.

$$Y_{it}^0 = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \varepsilon_{it}, \quad (4.21)$$

where:

- δ_t is a time-specific intercept common to all units,

- Z_i is a vector of observed covariates for unit i, with associated time-varying coefficients θ_t,
- λ_t is a vector of unobserved, time-varying factors affecting all units,
- μ_i is a vector of unit-specific, time-invariant characteristics,
- ε_{it} is an idiosyncratic error term.

With the linear factor model, the discussions in Abadie *et al.* (2010) and Abadie (2021) reveal the key assumption of identifying τ_{1t}. I interpret this assumption as a version of the **conditional exogeneity**. From the linear factor model in Equation 4.21, it is assumed that the error term ϵ_{it} is exogenous conditional on $\{Z_i, \mu_i, \lambda_t\}$. Therefore, SCM is subject to the bias from unobservables that vary across time periods and units. In this sense, it faces the same threat to identification as the DID. However, additional assumptions are needed for SCM, due to its objective of identifying the effect of individual treatment τ_{1t}. To see this, given the weights w_j, the difference between the true potential outcome of the treated unit Y_{1t}^0 and the synthetic potential outcome $\hat{Y}_{1t}^0$ is:

$$\begin{aligned} Y_{1t}^0 - \sum_{j=2}^{J+1} w_j Y_{jt}^0 = \theta_t \left(Z_1 - \sum_{j=2}^{J+1} w_j Z_j \right) \\ + \underbrace{\lambda_t \left(\mu_1 - \sum_{j=2}^{J+1} w_j \mu_j \right)}_{} \\ + \sum_{j=2}^{J+1} w_j (\varepsilon_{1t} - \varepsilon_{jt}). \end{aligned} \tag{4.22}$$

Given the **conditional exogeneity** assumption, the expectation of the difference is generally not equal to zero due to the second term, unless λ_t is zero (i.e., time homogeneity) or $\mu_1 - \sum_{j=2}^{J+1} w_j \mu_j$ is zero (i.e., $\mu \propto Z\theta_t$). However, neither of the conditions is probable in practice. Then, the question is under what conditions the second term becomes zero. Abadie *et al.* (2010) proves that a necessary condition is a sufficiently large number of pre-treatment periods T_0. However, this condition alone does not guarantee that the second term is zero. An implicit assumption is that the error term ε_{it} has finite moments, so that the second term

becomes zero as $T_0 \to \infty$. A *sufficient* condition that ensures that ε_{it} has finite moments is stationarity. Together, these two conditions are called **large T stationarity** (see Bottmer *et al.*, 2024). In summary, the SCM requires two fundamental assumptions: the conditional exogeneity and the large-T stationarity [6].

Lastly, to better understand these assumptions, I will discuss what ensures the SCM assumptions using an experimental design perspective from Bottmer *et al.* (2024). Using the idea of randomization of the experimental design, the **conditional exogeneity** assumptions is best ensured by the following two conditions:

1. **Random Assignment of Treatment**: The treated unit is selected randomly from the set of available units.
2. **Random Assignment of Treated Period**: The treated period is chosen randomly from the observed time periods, excluding early periods required for pre-treatment analysis.

If both the treatment unit and the treated period are randomly assigned, it is improbable that the treatment effect is subject to the bias of individual and time-varying unobservables. In other words, the conditional exogeneity is likely to hold. However, the plausibility of random assignments, especially that of the treated period, is less plausible in many real-world applications.

A reflection on SCM assumptions

In sum, researchers should interpret the treatment effect from SCM with extreme caution, due to the strong assumptions required for SCM to be valid. Moreover, both the assumption of **conditional exogeneity** and **stationarity** are inherently untestable, since the counterfactual outcome for the treated unit Y_{1t}^0 is not observed. In practice, it is generally recommended to have many time periods so that the plausibility of the assumptions can be examined with potential outcomes under control. The key idea of SCM has values and has been applied to improve methods such as DID in recent years. Interested readers are referred to

[6] There are also implicit assumptions that are needed for SCM. For example, the no spillover and no anticipation assumptions are excluded restrictions to rule out possible confounding effects

Arkhangelsky *et al.* (2021) and Xu (2017).

4.3 Summary

The general understanding of assumptions provides a valuable foundation for analyzing their role in stylized research designs. As discussed in the philosophy of science, assumptions are more than convenient simplifications; they shape the formulation of theories and influence how researchers interpret empirical evidence. In causal inference, this broader perspective helps us critically examine the assumptions that underpin various research methods, ensuring that they are not only methodologically sound, but also epistemologically justified. The structured approach to understanding assumptions through concepts, such as sufficiency and necessity, testability, and auxiliary assumptions, offers a framework to assess their credibility and limitations in empirical studies.

Applying these philosophical insights to stylized research designs reveals how assumptions serve as the backbone of causal identification strategies. Each method, whether instrumental variables, matching, difference in differences (DID), regression discontinuity design (RDD), or synthetic control methods (SCM), depends on a distinct set of assumptions that must be carefully evaluated to ensure valid causal conclusions. For example, in instrumental variable strategy, the assumptions of exogeneity, relevance, and exclusion restriction are sufficient to identify causal effects but are often difficult to validate empirically. The general understanding of assumptions highlights the importance of scrutinizing such premises and considering alternative sets of assumptions that might achieve the same identification goals with fewer constraints.

Similarly, DID analysis relies on the parallel trend assumption, which simplifies the complexity of treatment effects over time by assuming that, in the absence of treatment, the treatment and control groups would have followed similar trajectories. From a broader perspective, this assumption can be seen as a pragmatic compromise between theoretical abstraction and empirical feasibility. The insights from the general understanding of assumptions underscore the need to critically assess such simplifications and consider their plausibility in real-world settings.

Moreover, the Duhem-Quine thesis, which emphasizes the inter-

connectedness of assumptions, is particularly relevant when evaluating stylized research designs. Assumptions in causal inference methods rarely stand alone; they depend on auxiliary premises such as measurement accuracy, model specification, and data quality. Understanding this interplay helps researchers adopt a more holistic approach, combining robustness checks and sensitivity analysis to enhance the credibility of their findings.

In summary, the general understanding of assumptions provides a critical lens through which we can better appreciate and evaluate the assumptions embedded in stylized research designs. By drawing from philosophical insights, researchers can more effectively navigate the complexities of causal inference, ensuring that their assumptions are both justified and transparent.

5

Assessing Assumptions

Causal inference is based on a set of fundamental assumptions that enable researchers to draw valid conclusions about the effects of treatments or interventions. Due to the fundamental problem of causal inference, assumptions about potential outcomes cannot be directly tested using observed data. For example, the assumption of unconfoundedness is critical in many empirical settings but also highly controversial, due to the lack of direct evidence. This chapter examines the methodological challenges associated with assessing assumptions and introduces analytical tools designed to evaluate their plausibility. While no method can definitively establish or refute unconfoundedness, researchers can employ supporting analyses to enhance confidence in their causal estimates or identify potential biases.

The rigorous assessment of assumptions is essential to ensure the validity of causal inference methodologies, including regression adjustment, matching, and difference-in-differences (DID). If these assumptions are violated, causal effect estimates may be systematically biased, leading to incorrect conclusions. This chapter systematically explores two primary approaches for evaluating assumptions: sensitivity analysis and consistency tests. The sensitivity analysis evaluates how much violation

of assumptions would alter the estimated effects. If an unreasonable violation of assumptions is necessary to revert the conclusions, researchers can appeal to *reducio ad absurdum* and conclude that the assumptions are credible. Consistency tests, on the other hand, examine whether an identification strategy produces results that align with theoretically expected patterns in known contexts. The consistency between the estimation results and the prior expectations lends more credibility to the assumptions. Although these methods do not offer definitive proof of assumptions, they serve as crucial diagnostic tools to assess the reliability of causal claims.

5.1 Sensitivity Analysis

As illustrated by the example of estimating the price effect, for this estimation to be valid, one must assume that the price is unconfounded. In practice, this assumption is often implausible. Factors such as advertising, demand shocks, or product quality can simultaneously affect price and sales. Even if these variables are controlled for, it is still possible that other confounders exist. Sensitivity analysis can address this challenge by systematically "simulating" the possible confounding effect and examining how much the estimated price effect would change under different scenarios. If the analysis reveals that the confounding effect needs to be unrealistically large to rule out the observed price effect, one can appeal to *reductio ad absurdum* and argue that the assumption of unconfoundedness is credible. Therefore, the basic logic of the sensitivity analysis is to present hypothetical violations of assumptions to assess changes in causal conclusions.

This section systematically discusses the sensitivity analysis and covers its basic idea, logic, and structure. In addition, the common structure of different sensitivity analysis is presented to help researchers gain a better understanding. Selected methods for sensitivity analysis are summarized and discussed. Finally, the section concludes with practical challenges in implementing sensitivity analysis, such as interpreting subjective parameters, simulating unobserved confounders, and balancing identification with statistical power. Understanding these challenges is crucial for recognizing the limitations of current methods and guiding

future research to improve sensitivity analysis techniques.

5.1.1 The Fundamental Idea of Sensitivity Analysis

In empirical research that uses causal inference methods, sensitivity analysis is a fundamental tool used to assess the robustness of estimated effects to potential violations of the underlying assumptions. In causal inference, it helps researchers understand how sensitive their conclusions are to unobserved or imperfectly measured confounders that often bias the estimated effect of a treatment variable (D) and an outcome (Y). Causal inference methods, such as matching and difference-in-differences (DID), often rely on strong assumptions, such as no omitted variable bias or unconfoundedness, which may not hold in real-world applications. Sensitivity analysis provides a systematic way to explore how much the estimated effect could change under different scenarios where these assumptions are violated, offering a transparent assessment of the credibility of assumptions.

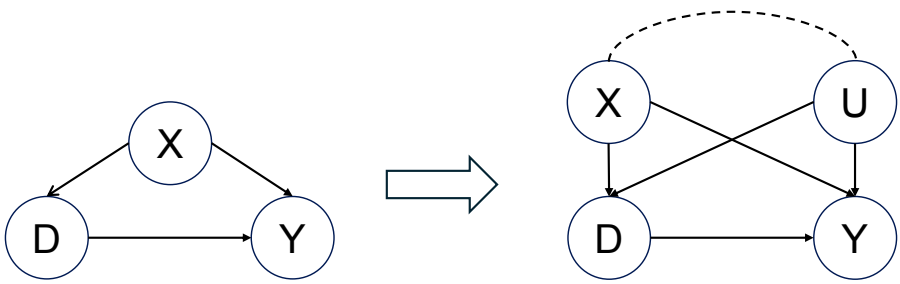
The underlying logic of sensitivity analysis is closely related to the philosophical concept of *reductio ad absurdum* - a form of argumentation that tests the validity of a claim by assuming the opposite and observing whether it leads to contradictions or implausible conclusions. In the context of sensitivity analysis, researchers begin by assuming that their assumptions are valid and then systematically introduce hypothetical violations (such as unobserved confounders) to examine the extent to which the causal estimation would be affected. If small or realistic deviations from the assumptions lead to implausibly large changes in the estimated effect, it suggests that the assumptions are not credible. In contrast, if the results remain stable under a broad range of plausible assumptions, it strengthens our confidence in the findings. In particular, if the causal estimation still holds even with an empirically or theoretically unreasonable violation, researchers can appeal to *reductio ad absurdum* and conclude that the assumptions are credible.

The Basic Logic of Sensitivity Analysis

Sensitivity analysis follows a *reductio ad absurdum* logic. The researchers assume that their causal assumptions hold and introduce hypothetical violations. If the causal estimation holds even under extreme violations, the credibility of the causal assumptions is reinforced.

A common structure underlying all sensitivity analyses can be illustrated using causal graphs. Consider two causal graphs side by side, as shown in Figure 5.1. The DAG on the left panel represents the general approach underlying many causal inference methods (e.g., matching and DID). The effect of D (treatment) on Y (outcome) is conditioned on a set of observed control variables X. The key assumption is the conditional exogeneity that the assignment of treatment is exogenous to the potential outcomes conditional on X. The DAG on the right panel introduces an additional unobserved confounder, U, which affects both D and Y. This is the tweaked DAG for the sensitivity analysis. In essence, researchers try to "simulate" the unobserved confounder U and introduce to the DAG to re-estimate the causal effect $P(Y \mid \text{do}(D))$. The reseadrchers can then compare the causal effect of the original DAG - $P^0(Y \mid \text{do}(D)) = \sum_x P(Y \mid D, X = x)$ with the tweaked DAG - $P^1(Y \mid \text{do}(D)) = \sum_{x,u} P(Y \mid D, X = x, U = u)$.

Figure 5.1: The Original vs. Tweaked DAG in Sensitivity Analysis



5.1.2 Selected Approaches to Sensitivity Analysis

Building on the foundational ideas presented in the previous section, I will discuss four specific methods for sensitivity analysis, each with unique strategies to address the challenges posed by unobserved confounding. These methods vary in how they model and simulate the unobservable confounder U and the specific contexts of the applications. Table 5.1 summarizes the methods. Moreover, this section is not a comprehensive review of all the methods for sensitivity analysis. Useful tools such as the E-value (VanderWeele and Ding, 2017) and the sensitivity analysis for mediation analysis (Imai *et al.*, 2010) are not covered. The methods can be understood with the common structure of the sensitivity analysis discussed in the previous section. For example, the E-value proposed by VanderWeele and Ding (2017) quantifies the minimum strength of association that U has with both treatment and outcome. Similarly, Imai *et al.* (2010)'s sensitivity analysis for mediation analysis models U as a mediator-specific confounder and vary the correlations between the error terms of the $M \sim D$ and $Y \sim D + M$ equation, where M denotes the mediator. Although these methods are not discussed in detail here, they illustrate the general principle that U is often simulated under specific assumptions. This section explores other prominent sensitivity analysis approaches, emphasizing their underlying assumptions and practical implications for causal inference.

Table 5.1: Summary of Sensitivity Analysis Methods

Method	**How U is simulated?**	**Context**
Rosenbaum	Shifts in the propensity score.	Matching
Oster	Proportional to selection of observables.	Regression
Blackwell	A flexible confounding function q.	General
Roth	Deviations in pre-treatment trends.	DID

Rosenbaum's sensitivity analysis

Rosenbaum's sensitivity analysis provides a formal framework to assess the design of observational studies using the matching method.

Researchers make the conditional exogeneity assumption in matching. However, the assumption is strong and there is always a concern that hidden confounders exist. Rosenbaum and Rubin (1983b) propose a method to evaluate the potential impact of an unobserved binary covariate, denoted U, on the treatment effect. In particular, the sensitivity analysis quantifies how much U need to influence the assignment of treatment D to (statistically) nullify the observed treatment effect. Therefore, Rosenbaum's sensitivity analysis "simulates" the confounder U based on the propensity score function $P(D \mid X)$.

The key element of Rosenbaum's sensitivity analysis is the introduction of the sensitivity parameter Γ, which measures the extent to which hidden biases might affect treatment assignment. Specifically, it considers a scenario in which subjects with identical observed covariates $(X_i = X_j)$ may differ in their odds of receiving the treatment due to unmeasured confounding variables. If there were no hidden bias, then the assignment of treatment would be purely random, which implies $\Gamma = 1$. However, if bias exists, the odds of receiving treatment for one individual could be up to Γ times greater than for another. Mathematically, Rosenbaum's model assumes that for matched pairs, the odds of treatment assignment for subject i relative to subject j satisfies the inequality:

$$\frac{P(D_i = 1 \mid X_i, U_i)}{P(D_j = 1 \mid X_j, U_j)} \leq \Gamma \tag{5.1}$$

For example, if $\Gamma = 2$, it suggests that an individual could be twice as likely to receive treatment due to unmeasured factors.

A central feature of Rosenbaum's method is the computation of bounds on quantities for statistical inference, such as p-values of the treatment effect. By varying the level of confoundedness or Γ, researchers can know how sensitive the statistics of interest are to the level Γ. For a fixed value Γ, the sensitivity analysis produces an interval of possible p-values rather than a single value. From the interval, researchers can learn the lower and upper bounds of the p values. These bounds increase as the bias Γ increases, making the results more uncertain.

The following example illustrates the sensitivity analysis in a matched pair design, where people are paired based on similar characteristics or covariates to control for potential confounding. Suppose that there are

S pairs, $s = 1, \ldots, S$, of two subjects, one treated, one control, matched for observed covariates. The researcher wants to test the null hypothesis of no treatment effect, that each subject would have the same response under treatment or control or $\Delta Y_s = Y(s = 1) - Y(s = 0) = 0$. The absolute differences, $|\Delta Y_s|$, range from 1 to S, and Wilcoxon's signed rank statistic, W, is the sum of the positive differences ranks, $\Delta Y_s > 0$.

Under the null hypothesis of no effect, W is approximately normally distributed with a moderately large S:

$$\begin{cases} \mu &= \dfrac{S(S+1)}{4} \\ \sigma^2 &= \dfrac{S(S+1)(2S+1)}{24} \end{cases} \tag{5.2}$$

Without bias or $\Gamma = 1$, for a matched pair, the probability of being treated is equal to one-half. However, for $\Gamma > 1$, the odds of receiving treatment for one subject compared to another can be up to Γ times greater. This means that for each matched pair, the probability of being treated lies within $1/(1+\Gamma)$ and $\Gamma/(1+\Gamma)$. The distribution of W is approximately bounded between two Normal distributions, with expectations:

$$\mu_{\max} = \lambda \frac{S(S+1)}{2}, \quad \mu_{\min} = (1-\lambda)\frac{S(S+1)}{2} \tag{5.3}$$

and variance:

$$\sigma_\Gamma^2 = \lambda(1-\lambda)\frac{S(S+1)(2S+1)}{6} \tag{5.4}$$

where $\lambda = \Gamma/(1+\Gamma)$. With the two bounding distributions, the upper and lower bounds of the p values of the test statistic W can be calculated using the Z-test.

For the model-based approach to Rosenbaum's sensitivity analysis, refer to Rosenbaum and Rubin (1983b) and Imbens and Rubin (2015). Although Rosenbaum's sensitivity analysis is widely adopted in various fields (e.g., Chang *et al.*, 2023; Kim *et al.*, 2015, in marketing), it is limited to analyses that use the matching method. Next, I will discuss an approach applicable to a general linear regression model under selection bias.

Oster's selection bias approach

A new method for conducting sensitivity analysis to assess selection bias, as proposed by Oster (2019), has been widely adopted in different disciplines (e.g., Hassan *et al.*, 2023; Wang and Goldfarb, 2017, in marketing). This approach is favored due to its broad applicability to regression models, simple implementation, and ease of interpretation. The method revolves around two key concepts: *coefficient stability* and *R-squared movements*, which together offer insight into how much omitted variables could bias the estimated treatment effect.

The method starts with a common approach in empirical research to assess omitted variable bias: to examine how much the coefficient of the treatment variable changes when additional controls are added to the regression model. If the coefficient remains stable (i.e., does not change significantly), researchers often conclude that omitted variable bias is limited. However, *coefficient stability* alone is not enough to draw such conclusions. One important reason is that a coefficient might appear stable simply because the added controls explain very little of the outcome variance, rather than because they adequately capture potential confounders.

Based on this observation, Oster (2019) argues that *coefficient stability* must be interpreted in conjunction with how much the controls increase the explanatory power (R-squared) of the model. If R^2 increases significantly after adding controls, it suggests that the observed variables capture meaningful variation in the outcome, increasing the confidence that they can also capture the influence of omitted confounders. If R^2 barely changes, even when controls are added, it implies that the observed variables do not explain much additional variance, making it likely that omitted variables could still have a large impact.

Based on the aforementioned intuitions, Oster (2019) formalize the idea using the following approach:

$$\hat{\beta} = \hat{\beta}_{\text{uncontrolled}} - \delta \left(\frac{R^2_{\text{max}} - R^2_{\text{controlled}}}{R^2_{\text{controlled}} - R^2_{\text{uncontrolled}}} \right) \cdot \left(\hat{\beta}_{\text{uncontrolled}} - \hat{\beta}_{\text{controlled}} \right), \quad (5.5)$$

where $\hat{\beta}_{\text{controlled}}$ and $\hat{\beta}_{\text{uncontrolled}}$ are the coefficients of the treatment

variable from the regression with and without control variables. The same subscriptions for R^2 and R^2_{max} is the maximum obtainable R^2 (e.g., $R^2_{\text{max}} = 1$). δ is the key sensitivity parameter, which quantifies the relative importance of selection on unobservables compared to observables:

$$\delta = \frac{\text{Selection on Unobservables}}{\text{Selection on Observables}} \tag{5.6}$$

If δ is assumed to be 1, it implies that the omitted variables are no more strongly related to the treatment than the observed variables. If results remain robust under this assumption, it strengthens confidence in the findings. If higher values of δ are needed to explain away the effect, it suggests greater robustness.

However, it should be noted that the interpretation of δ is only valid under a strong assumption - the proportional selection on the treatment assumption [1]. This assumption states that the degree of selection on unobservables can be approximated by the degree of selection on observables. It makes the sensitivity parameter δ interpretable. The implication is that unobservables or confounding variables are "simulated" based on the propensity score function $P(D \mid X)$. However, if unobserved variables are fundamentally different in their relationship to treatment than observed variables, the sensitivity analysis might understate or overstate the bias of the omitted variable. Formally, the requirement is the two conditional probabilities $P(D \mid X)$ and $P(D \mid U)$ are statistically close:

$$\Delta(X, U) \propto \sum_{\alpha} \left| \delta P(D \mid X = \alpha) - P(D \mid U = \alpha) \right|, \tag{5.7}$$

with $\Delta(X, U)$ a negligible function. Lastly, the assumption implies that the sensitivity analysis can *only* test the robustness against the existence of particular unobservables which satisfy Equation 5.7.

Blackwell's selection bias approach

Although the selection bias approach in Oster (2019) has gained much attention, one limitation is this approach is based closely on linear

[1] Other assumptions are also required to obtain the exact formula in Equation 5.5. Interested readers can refer to Oster (2019) for more details.

regression. In recent years, researchers increasingly combine non-linear and non-parametric models into causal inference. For example, Wager and Athey (2018) develop the causal random forest to learn heterogeneous treatment effect from randomized experiments and later extend its application to observational data (Athey *et al.*, 2019). Researchers also try to apply machine learning models to improve matching designs (e.g., Ferri-García and Rueda, 2020). Therefore, in these cases, many current sensitivity analyses cannot apply because they often assume linear and parametric functional forms. To this end, Blackwell (2014) develops a general sensitivity analysis that can be applied in more complex settings.

Blackwell's method provides a practical approach to assessing omitted variable bias by directly focusing on the consequence of selection bias in treatment assignment rather than making assumptions about unknown confounders. By introducing a confounding function, researchers can systematically adjust the level of bias and evaluate its impact on causal estimates, offering a structured way to explore deviations from assumptions like conditional exogeneity. A key advantage of this approach is its flexibility, as it integrates easily with common causal inference methods such as regression, matching, and weighting without requiring complex derivations. In addition, the method provides a clear quantitative measure of sensitivity to selection bias, helping researchers assess the robustness of their findings under plausible violations of exogeneity.

Blackwell's method for sensitivity analysis is mathematically grounded in the concept of a confounding function, which quantifies the extent of selection bias in treatment assignment. The confounding function, under a binary treatment D, is mathematically defined as,

$$q(d, X) = \mathrm{E}(Y^d \mid D = d, X) - \mathrm{E}(Y^d \mid D = 1 - d, X) \tag{5.8}$$

If no selection bias, then $q(d, X) = 0$, as the potential outcomes are conditional exogenous with respect to the treatment. An intuitive interpretation of the bias function is that it captures the selection bias. For example, under Manski (1997)'s monotone treatment selection assumption and assuming the treatment always increases potential outcomes, the bias function $q(1, X)$ should be always positive. Next, we can obtain

the expected potential outcomes with the confounding function,

$$\begin{cases} \mathrm{E}\left(Y^1 \mid D=1, X\right) & = \mathrm{E}\left(Y^1 \mid D=0, X\right) + q(1, X) \\ \mathrm{E}\left(Y^0 \mid D=0, X\right) & = \mathrm{E}\left(Y^0 \mid D=1, X\right) - q(0, X) \end{cases} \tag{5.9}$$

From these equations, the treatment effect can be derived, given the bias function $q(d, X)$.

Moreover, as the key element of the approach, Blackwell (2014) proposes adjusting the observed outcome Y_i to account for potential bias. The adjusted outcomes are then obtained by accounting for potential unmeasured confounding with the confounding function. The adjustment is achieved by modifying the observed outcome Y_i to remove the estimated bias introduced by the selection effects. The adjusted outcome is expressed as:

$$Y_i^q = Y_i - q(D_i, X_i) \cdot (1 - D_i)e(X_i) + D_i(1 - e(X_i)), \tag{5.10}$$

where $e(X_i)$ is the propensity score, defined as $P(D_i = 1 \mid X_i)$[2]. The key decision then remains how to select a reasonable bias function. The selection is often not straightforward as the function is totally unbounded, as researchers do not observe the confounding variables. Blackwell (2014) suggests using a simple yet flexible parameterization, such as $q(d, x; \gamma) = \gamma(2d - 1)$, which assumes that the treated units are inherently better or worse off compared to the control units depending on the sign of γ. This approach is easy to implement but makes a similar assumption as in Oster (2019) that the selection of observables and unobservables is proportional. However, it is difficult to specify the bias as a function of observables and treatment unless researchers have domain knowledge and theoretical expectations about the unobservables.

Given the bias function $q(d, X)$, researchers can follow the procedure below to perform the sensitivity analysis.

1. Estimate the treatment effect with a model of $Y \sim f(D, X)$ and obtain the propensity score function $e(X) = P(D \mid X)$.
2. Select and vary the level of bias by setting a range of key parameters. For example, the value of γ can be varied if $q(d, X) = \gamma(2d - 1)$.

[2]For the detailed derivation, please see p. 173 of Blackwell (2014).

3. Obtain the "debiased" outcome Y_i^q using Equation 5.10.
4. With Y_i^q, re-estimated the treatment effect with the original model $Y \sim f(D, X)$, but replace Y with Y_i^q.
5. Collect estimated treatment effects and examine critical levels of bias that nullify treatment effects.

The procedure is flexible, as any models $f(D, X)$ and $e(X)$ can be nested with the procedure. For example, $f(D, X)$ and $e(X)$ can be obtained by applying the causal random forest or any other causal machine learning approach.

Sensitivity analysis for DID

In this section, I will review the sensitivity analysis approach for DID proposed in Rambachan and Roth (2023). The initial motivation behind the method is the common diagnostics in DID analysis, where tests of pre-trends are used to justify the parallel trends assumption stems. The fundamental idea is that if the treatment and control groups exhibited similar trends before treatment was introduced, it is reasonable to assume that they would have continued to do so in the absence of treatment. However, the similarity in the pre-trend is neither necessary nor sufficient for the parallel trend assumption. The best one can conclude from statistical evidence supporting the similarities in pretrends is that the parallel trend assumption is more plausible.

Recent research (e.g., Freyaldenhoven *et al.*, 2019; Roth, 2022) has highlighted several limitations of this approach, such as the low power of pre-trends tests to detect meaningful differences and the statistical issues arising from conditioning on passing such tests. Recognizing these challenges, Roth *et al.* (2023) proposes a more flexible framework that leverages pre-trends not as a binary test but as informative bounds on the deviations of post-treament potential outcomes from the parallel trend assumption.

I will use the same notation and focus on binary treatment (treated vs. control) in two periods (before vs. after). Potential outcomes are denoted as $Y_{it}(D)$ or the potential outcome of unit i in the treatment state D at time period t. Under the parallel trend assumption, the

pre-trend is equal to the post-trend of controlled potential outcomes.

$$\begin{aligned}&\mathrm{E}(Y_{i2}(0) \mid D_{i2}=1)-\mathrm{E}(Y_{i2}(0) \mid D_{i2}=0) \\ &\quad=\underbrace{\mathrm{E}(Y_{i1}(0) \mid D_{i2}=1)-\mathrm{E}(Y_{i1}(0) \mid D_{i2}=0)}_{\text{Pre-trend}}\end{aligned} \tag{5.11}$$

Denote the pre-trend and post-trend as δ_1 and δ_2, respectively. Suppose the parallel trend assumption does not hold, and the pre- and post-trend differ by Δ,

$$\delta_2 = \delta_1 + \Delta \tag{5.12}$$

The bias in the estimated treatment effect (ATT) is therefore Δ as shown below,

$$\begin{aligned}\mathrm{ATT} &= \mathrm{E}\left(Y_{i2}\left(1\right) \mid D_{i2}=1\right)-\mathrm{E}\left(Y_{i2}\left(0\right) \mid D_{i2}=1\right) \\ &= \underbrace{\left(\mathrm{E}\left(Y_{i2}\left(1\right) \mid D_{i2}=1\right)-\mathrm{E}\left(Y_{i2}\left(0\right) \mid D_{i2}=0\right)-\delta_1\right)}_{\widehat{\mathrm{ATT}}}-\Delta\end{aligned} \tag{5.13}$$

With the above derivation, if the bias value Δ is specified, we can re-estimate the model with the known pre- and post-trend difference. For example, for a simple two-period DID, we can non-paramterically estimate the ATT and bootstrap the confidence interval under different levels of Δ. Then we can compare the original estimate under the parallel trend assumption $\Delta = 0$ and those with $\Delta \neq 0$, and examine when the original estimate is nullified or what level of bias leads to significantly different estimates. Roth *et al.* (2023) propose a partial identification approach by assuming that the bias belongs to an interval with $\Delta \in \{\underline{\Delta}, \overline{\Delta}\}$. Given this assumption, the true ATT belongs to $\{\widehat{ATT} - \overline{\Delta}, \widehat{ATT} - \underline{\Delta}\}$. Statistical inference of the probability that the true ATT belongs to the set can be made using the moment inequality-based inference or the fixed-length confidence interval approach.

The above derivation seems straightforward, but the bias term Δ is inherently difficult to specify because, in principle, it can take arbitrarily large values. To choose a reasonable and defensible value for empirical applications, Roth *et al.* (2023) propose defining Δ relative to observed pre-trends such that $|\Delta| \leq \lambda |\delta_1|$, where λ is the sensitivity parameter and a larger λ leads to a wider range of bias. This approach aligns with

the intuition behind traditional pre-trend tests. The underlying assumption here is that the magnitude and direction of any post-treatment bias should be similar to the discrepancies observed in the pre-treatment period.

However, this approach is not without its limitations. By assuming that post-treatment bias is proportional to pre-treatment trends, it imposes a specific structure on the bias that may not capture more complex or dynamic sources of confounding. This limitation is similar to the challenges faced in other sensitivity analysis frameworks, such as the Blackwell method, where the choice of the bias function is similarly based on assumed relationships with observable data.

5.1.3 Practical Challenges to Sensitivity Analysis

Sensitivity analysis is a valuable tool for assessing the credibility of the assumptions underlying causal inference. By systematically "challenging" assumptions and quantifying their impact on the results, sensitivity analysis enables researchers to explore the robustness of their conclusions in the presence of potential violations of key assumptions. However, while it is undeniably useful, sensitivity analysis is not without practical challenges. These challenges arise from the subjective nature of parameter interpretation, the reliance on specific simulation frameworks for unobservables, and the interaction between identification and statistical power. Understanding these limitations is crucial to ensure that sensitivity analysis is used effectively and interpreted appropriately.

One of the most pressing challenges lies in the interpretation and benchmarking of the sensitivity parameters. These parameters quantify the degree of deviation from key assumptions, but are often difficult to interpret in a meaningful way. A case in point is the sensitivity analysis for mediation analysis developed by Imai *et al.* (2010). The sensitivity parameter is the correlation between the mediator and the outcome produced by the confounders. If a researcher finds a critical value of the correlation equal to 0.5 that nullifies the mediation effect, the question remains: *Is it an indication of the credibility or unreliability of the mediation analysis?* In practice, researchers must decide what constitutes a deviation "large" or "small", and these judgments are

typically based on subjective beliefs and domain-specific knowledge. This subjectivity can lead to disagreements among researchers, as the chosen benchmarks may reflect varying perspectives and expectations. Without clear and universally accepted guidelines for what constitutes reasonable parameter values, the conclusions of sensitivity analysis may lack consistency and comparability between studies.

Another critical challenge is how the unobservable confounder U is simulated within sensitivity analysis frameworks. As highlighted in the previous section, many methods rely on specific assumptions about U, such as its relationship to observed covariates or its proportional impact on the outcome. While these approaches provide a structured way to evaluate robustness, they also introduce a risk of bias. Simulating U based on specific assumptions can inadvertently align the results with those assumptions, leading to false positives in evaluating the credibility of causal inference. By modeling U in constrained ways, sensitivity analysis can overlook more complex or non-linear forms of confounding, which may distort its conclusions.

Finally, sensitivity analysis often conflates identification issues with challenges related to statistical power. The methods require the use of a specific data sample, and the robustness of the results depends on the sample's size and variability. In cases of low statistical power, sensitivity analysis can produce misleading results, such as failing to detect meaningful deviations from assumptions or exaggerating the uncertainty of causal estimates. This conflation means that even theoretically sound sensitivity analyses can yield false negatives when the sample does not provide sufficient information to assess assumptions effectively. Researchers must carefully disentangle issues of statistical power from the validity of the assumptions being tested to avoid over-interpreting sensitivity analysis results.

5.2 Consistency Tests

When estimating a causal effect, researchers often begin with very little direct knowledge of its true magnitude or nature - indeed, the very reason for undertaking causal estimation is to uncover these unknowns. However, in order to develop a credible identification strategy or research

design, as discussed in Chapter 3, researchers must impose strong assumptions, such as unconfoundedness. The power of an identification strategy therefore depends on the credibility of these assumptions, but their validity is typically unobservable. This creates a fundamental tension. Although the identification strategy serves as a measurement tool for the causal effect, we do not have an independent way to verify its validity for the particular treatment-outcome relationship under study.

In addition to sensitivity analysis, another possible avenue for assessing assumptions is consistency tests. In consistency tests, researchers evaluate whether the assumptions underlying the identification strategy hold in contexts where we have reasoned expectations about the treatment effect. Even though we do not know the true causal effect, we often have theoretical insights, stylized facts, or prior empirical findings that suggest how the treatment effect should behave under specific conditions. For example, a well-reasoned expectation may be that a treatment should have no effect on a subgroup that is theoretically immune to it or that a treatment should not alter past outcomes. If our identification strategy is valid, then applying it in these specific contexts should yield results that align with our prior expectations. When results systematically deviate from these expectations, it raises doubts about the validity of the identification strategy and the credibility of assumptions.

5.2.1 Basics of Consistency Tests

To better understand consistency tests, let us start with a popular form of consistency tests is known as "placebo tests" (Eggers *et al.*, 2024). In placebo tests, researchers have prior beliefs about the lack of treatment effects in known scenarios. The test estimates pseudo-causal effects under conditions where the true effect is expected to be zero. Researchers can then observe the consistency between their prior beliefs and actual estimation of the pseudo-causal effects. Placebo tests can generally be categorized into three types: pseudo-outcome tests, pseudo-treatment tests, and pseudo-group tests (see Imbens and Rubin, 2015, Chapter 21). They differ in how the lack of treatment effect originates.

Pseudo-outcome placebo tests involve selecting a variable that is known to be unaffected by treatment and using it as a "placebo" outcome. By the same logic, the *pseudo-treatment placebo test* assigns a pseudo-treatment, a variable that is known to have no causal effect on the outcome. Lastly, the *pseudo-group placebo test* assesses a subgroup of individuals who, according to theory or institutional knowledge, should not be affected by treatment.

Placebo tests are an "extreme" form of consistency tests. In general, researchers should have prior beliefs about the effect of treatment in certain scenarios and expect the identification strategy and assumptions to produce consistent predictions when applied to these scenarios. Hence, the consistency test is defined as follows (Definition 5.1).

Definition 5.1 (Consistency Tests)**.** A consistency test is a diagnostic tool used in causal inference to assess the plausibility of an identification strategy and assumptions by evaluating whether estimated causal effects align with expected patterns in specific scenarios where the true effect is known or can be reasonably inferred.

Having a formal definition of consistency tests, the next fundamental question is: *What are consistency tests actually testing?* In essence, the reasoning behind consistency tests is in alignment with the principles of plausibility reasoning, as discussed in probability theory (Jaynes, 2003, Chapter 1). By definition, plausibility reasoning is a structured way of making inferences in situations where we lack complete information, but must nevertheless draw conclusions based on logical consistency and prior knowledge. Unlike deductive reasoning, which leads to fixed conclusions given true premises, plausible reasoning is probabilistic and dependent on context. Using plausibility reasoning, one can assess whether new evidence strengthens or weakens an existing belief, making it an essential tool for evaluating the credibility of assumptions in causal inference.

The fundamental principle of plausibility reasoning is that conclusions drawn from incomplete information should be consistent with what we already know. The reasoning process relies on *weak syllogisms*, which do not establish certainty but increase or decrease the plausibility of a claim. As a general example, assume that we have two statements

or assumptions A and B. A standard deductive reasoning works with the following strong syllogism,

$$\frac{\begin{array}{c}\text{If } A \text{ is true, then } B \text{ is true.}\\ A \text{ is true.}\end{array}}{\text{Conclusion: } B \text{ is true.}} \tag{5.14}$$

However, in almost all the situations that require statistical decision-making, we do not have the right or complete information to allow for the standard deductive reasoning. Instead, we fall back on the *weaker syllogisms*:

$$\frac{\begin{array}{c}\text{If } A \text{ is true, then } B \text{ is true.}\\ B \text{ is true (false).}\end{array}}{\text{Conclusion: } A \text{ becomes more (less) plausible.}} \tag{5.15}$$

In the context of causal inference, consistency tests function similarly: they do not definitively prove the validity of identification assumptions, but offer indirect evidence about whether it is likely to be valid.

From a formal perspective, consistency tests in causal inference can be formulated as a plausibility reasoning process as follows. In consistency tests, researchers aim to assess the plausibility of the *fundamental assumption* A_0 (e.g., unconfoundedness). We have some prior belief (assumption) about the effect of the treatment given A_0. The prior belief often states another condition or *auxiliary assumption* A_1, where the treatment effect should be like what is described in the statement B. For example, under the assumption of unconfoundedness (condition A_0), treatment should not have an effect on the outcome before treatment (condition A_1). Therefore, we should not find any treatment effect (conclusion B). Formally, consistency tests assume the following weak syllogism:

$$\frac{\begin{array}{c}\text{If } A_0 \text{ is true, then } \overline{A_1} \vee B \text{ is true.}\\ \overline{A_1} \vee B \text{ is true (false).}\end{array}}{\text{Conclusion: } A_0 \text{ becomes more (less) plausible.}} \tag{5.16}$$

In Equation 5.16, $\overline{A_1}$ is read as "not A_1; $\overline{A_1} \vee B$ is the OR operation in Boolean algebra and stands for $A_1 \rightarrow B$ (A_1 implies B) or the statement "if A_1 is true, then B is true." Thus, different consistency

tests can be seen as instantiations of this general framework, where researchers choose different A_1 and examine if the statistical conclusions are consistent with B. The consistency increases the plausibility of A_0.

5.2.2 Examples of Consistency Tests

In the next section, I will explore two examples of consistency tests in empirical research and demonstrate how they align with the proposed general framework. The examples are selected to illustrate the widespread use of consistency tests in different studies, rather than to serve as special or representative cases. Many other applications exist, but these examples provide a brief glimpse into how consistency tests are applied in practice, offering insight into their role in evaluating the credibility of causal inference.

The first example is about lending more credibility to mediation analysis with consistency tests in consumer psychology research. Longoni *et al.* (2019) investigates consumer reluctance to accept AI-based healthcare services. They study consumer receptivity to AI in medical settings, including willingness to utilize AI healthcare, willingness to pay, and evaluation of AI provider performance. The study identifies uniqueness neglect as the key mediator, which refers to consumers' concern that AI providers cannot account for their unique characteristics, circumstances, and symptoms as well as human providers. In Study 6, the authors test whether uniqueness neglect mediates the effect of provider type (human vs. AI) on consumers' likelihood to follow a medical recommendation.

The identification of mediation effect relies on the critical assumption that the mediator is unconfounded, which often does not hold in behavioral experiments where the mediator is measured after the main task rather than manipulated. In this research, the fundamental assumption (A_0) is that the mediation effect in Study 6 is estimated without bias. To further strengthen confidence in this mediation effect, Study 7 was designed to weaken uniqueness neglect by introducing personalization as a treatment. Here, the auxiliary assumption (A_1) is that personalization effectively reduces uniqueness neglect. A moderation analysis in Study 7 showed that, consistent with expectations,

when AI care was framed as personalized, resistance to AI healthcare disappeared (B).

The second example is a reduced-form analysis that investigates how changes in household income and wealth influence the demand for private-label (store-brand) products (Dubé *et al.*, 2018). To establish causality, the authors exploited the geographic variation in the severity of the economic shocks of the Great Recession on income and wealth. Their empirical strategy uses household panel data from Nielsen's Homescan database, matched with store-level scanner data and Zillow's local housing price indices to measure wealth effects. The key assumption (A_0) is that within-household changes in income and wealth are "as good as randomly assigned" once other factors are controlled for (i.e., conditional exogeneity). This assumption ensures that their estimates capture the causal effect of income and wealth on private-label demand, rather than being driven by unobserved household characteristics or pre-existing trends.

To further examine the plausibility of their assumption, they used pre-trends analysis as a placebo test. The central idea is that if the key assumption (A_0) holds that within-household changes in income and wealth during the Great Recession are exogenous, then the pre-recession trends in private-label demand should be independent of the income and wealth during the great recession (A_1). Therefore, changes in the private-label shares for households and geographic areas (e.g., ZIP codes or counties) should not be correlated with the changes in income, wealth and unemployment during the recession (B). In the paper, the authors find that the correlations are very small (close to zero) and mostly not statistically significant, increasing the plausibility of their assumptions.

5.2.3 Determining the Power of Consistency Tests

Consistency tests provide an essential layer of scrutiny in causal inference, ensuring that the estimated effects align with theoretical expectations under specific conditions. By grounding them in the logic of *plausibility reasoning*, we can see them as structured evaluations of whether an identification strategy behaves in a logically coherent manner. The examples in the previous section show how consistency

tests provide evidence that strengthens researchers' confidence in the empirical strategies and the underlying assumptions.

Despite their usefulness in assessing the credibility of causal inference strategies, consistency tests raise several open questions that require further theoretical and methodological development. The key challenge concerns the power of consistency tests. We often have intuitive judgments about which tests provide stronger evidence against the credibility of assumptions in research designs. However, there is no formal framework that clearly explains what determines the strength of a test. For example, finding a significant estimated effect in a pre-treatment period is widely regarded as strong evidence against the unconfoundedness assumption, as it directly contradicts the fundamental causal structure. In contrast, consistency tests based on treatment effects in certain subpopulations may be perceived as weaker, as they depend more on auxiliary theoretical assumptions about treatment heterogeneity. Thus, an important unresolved question is how to systematically quantify the evidentiary value of different consistency tests.

In the previous section, we have established that consistency tests are essentially plausibility reasoning. Although qualitative reasoning proves to be valuable, in empirical work it is fundamentally important to assign some degree of plausibility to consistency tests. Through quantification, researchers can know the importance of different factors that determine plausibility. Quantification builds the foundation for comparing and synthesizing between different consistency tests. At a fundamental level, I follow the *logic of plausibility reasoning*, as formalized by the Cox theorem and Jaynes's probability desiderata (see Jaynes, 2003, p.17). Following the basic desiderata, one can project the logical statement onto the probability space.

Next, the basic logic statement for consistency tests is $A_0 \to (A_1 \to B)$. This logical statement can be further expressed as an equivalent statement using Boolean algebra,

$$\begin{aligned} A_0 \to (A_1 \to B) &= A_0 \to \left(\overline{A_1} \vee B\right) = \overline{A_0} \vee \overline{A_1} \vee B \\ &= \overline{A_0 A_1} \vee B = (A_0 \wedge A_1) \to B \end{aligned} \tag{5.17}$$

In the above equation, $\vee$ and $\wedge$ represent the *OR* and *AND* operations. The equation is obtained by first applying the rule that $A \to B$ is

represented as $\overline{A} \vee B$ in Boolean algebra. Then, associative law and de Morgan's Theorem are applied to the third and fourth equality. For simplicity, let us first denote a "complex assumption" $A = A_0 \wedge A_1$. We will further discuss this later. In addition, for the clarity of discussion, here we ignore the background information which is contained in the available to researchers. All conditional probabilities can be expanded by further conditioning on the data, e.g., $P(A \mid B, \text{Data})$, and all conclusions remain valid. Under the weak syllogism as in Equation 5.16, the probability rules based on Jaynes's probability desiderata show that (i.e., the Bayes rule),

$$P(A \mid B) = P(A) \frac{P(B \mid A)}{P(B)} \tag{5.18}$$

Based on Equation 5.18, we first observe that under the logical statement behind the weak syllogism, $P(B \mid A) = 1$. Since $P(B) \leq 1$, we must have $P(A \mid B) > P(A)$, which corresponds to the weak syllogism that if B is true, A becomes more plausible. We can also answer the basic question: What determines whether the evidence in the form of the consistency of prior belief in B elevates the plausibility of A? The answer of the equation is that since $P(B \mid A)$ cannot be greater than unity, a large increase in plausibility of A conditional on B being true can only occur when $P(B)$ is small. The implication is that the plausibility of B is low in nature or in a general case. On the other hand, if knowing that A is true can only make a negligible increase in the plausibility of B (small $P(B \mid A)$), then observing B can in turn make only a negligible increase in the plausibility of A. The implication here is that B are much more likely to be true in a specific situation when A is true.

To illustrate the above conclusions, I will use *the balance test* commonly used in the matching method as an example. In balance tests, researchers believe that if the covariates are balanced (B), then the assumption that treatment is randomly assigned (A) is more plausible. The power of the balance test depends on two conditions: 1) $P(B)$ or the likelihood that the covariates are balanced in a general setting (the treatment is not necessarily randomly assigned) and 2) $P(B \mid A)$ or the likelihood of covariates being balanced in the setting where the

treatment is randomly assigned. A balance test only increases our confidence in A if $P(B \mid A) > P(B)$, which implies $P(A \mid B) > P(A)$ (A becomes more plausible given that B is true). In other words, if covariates are much more likely to be balanced under random assignment, balance tests provide evidence for the random assignment or exogeneity assumption.

However, notice that A is the joint event of A_0 and A_1. The observation that B is true increases the plausibility of A_0 and A_1 are both true. Our end goal here is to find evidence to enhance the plausibility of A_0 (i.e., $P(A_0 \mid B) > P(A_0)$). Therefore, let us expand on Equation 5.18.

$$P(A_0 \mid B) = \left[\frac{P(A_1 \mid A_0)}{P(A_1 \mid A_0 B)} P(A_0)\right] \frac{P(B \mid A)}{P(B)} \tag{5.19}$$

From the equation, given $P(B \mid A) > P(B)$, a sufficient condition for $P(A_0 \mid B) > P(A_0)$ is $P(A_1 \mid A_0) > P(A_1 \mid A_0 B)$. On the contrary, if $P(A_1 \mid A_0) < P(A_1 \mid A_0 B)$, we may have $P(A_0 \mid B) < P(A_0)$, even if $P(B \mid A) > P(B)$. In practice, the best case is probably to find a condition A_1 that is always true conditional on A_0, regardless of whether B is true. Next, I will show how we can use this condition to explain why a placebo test based on A_1, which asserts that there is no treatment effect in the pre-treatment period, is generally considered more powerful than a placebo test where A_1 assumes no treatment effect within a specific subgroup.

The key distinction between these two placebo tests lies in the nature of these assumptions. In the case of the pre-treatment placebo test, A_1 states that *no treatment effect should be observed before treatment occurs*, which is almost a stylized fact, as causality cannot operate backward in time. Under this condition, we can reasonably assume that $P(A_1 \mid A_0) = P(A_1 \mid A_0 B)$, since the truth value of B is irrelevant to the plausibility of A_1.

In contrast, for a subgroup-based placebo test, A_1 typically states that *a particular subgroup of individuals should not be affected by treatment*. However, our confidence in this assumption is generally weaker compared to the pre-treatment placebo test. Unlike the time-based assumption, which is grounded in an immutable temporal constraint, the assumption that certain individuals are unaffected by the treatment

is more speculative and context-dependent. Furthermore, maintaining the condition $P(A_1 \mid A_0) > P(A_1 \mid A_0 B)$ is challenging in the subgroup placebo test. This is because a positive finding in B (evidence of no effect in the subgroup) would *reinforce rather than undermine* the plausibility of A_1.

Determinants of The Power of Consistency Tests

Researchers can construct a more powerful consistency test by finding 1) a conclusion B that is more likely to be true only if both the fundamental assumption A_0 and the auxiliary assumption A_1 are true; and 2) an auxiliary assumption A_1 whose plausibility does not depend on whether the conclusion B is true.

5.2.4 Open Questions

Although Proposition 5.2 outlines the general conditions for a consistency test to be more powerful, there are still many open questions. For example, how to aggregate evidence from multiple consistency tests within the same research design and how to compare consistency tests across studies. It is important to address these open questions in future research to better apply consistency tests in empirical work.

One of the key challenges is that researchers often conduct several consistency tests to assess the plausibility of their identification strategy, but there is currently no well-defined quantitative framework for synthesizing their findings into a unified evaluation. In some cases, different consistency tests may yield conflicting results. Some tests may reinforce confidence in the validity of the assumptions, while others may raise doubts. This creates a challenge in weighing and interpreting the overall evidence: should a single failed consistency test be sufficient to question the entire identification strategy, or can it be outweighed by other tests that support its credibility?

In addition, different tests may have varying degrees of informativeness, and their relative weight in the overall assessment remains an open methodological question. Without a formal approach to aggregating these results, researchers must rely on *ad hoc* reasoning, which may

introduce subjective biases. Developing a quantitative framework for evaluating and combining the results of multiple consistency tests would improve the rigor of causal inference, providing a systematic way to assess the strength of evidence for or against the validity of a research design. Such a framework could incorporate probabilistic reasoning, Bayesian updating, or statistical measures of test informativeness to offer a more objective and transparent approach to evaluating the credibility of causal assumptions.

Another challenge is the comparability of consistency tests across different studies. Since empirical applications construct consistency tests based on theoretical predictions and domain knowledge specific to their settings, there is no standardized way to compare them across different empirical contexts. This makes it difficult to generalize insights from one study to another or to establish best practices for designing consistency tests.

5.3 Conclusion

In this chapter, I have discussed two approaches in our toolkit for assessing assumptions. Sensitivity analysis systematically introduces hypothetical violations of key assumptions and examines the impact on estimated effects. By assessing how much the estimated effect changes under different levels of assumed confounding, researchers can gauge the plausibility of their assumptions. The fundamental logic of the sensitivity analysis follows a *reducio ad absurdum* approach. If the estimated effect remains consistent even given "unreasonable" violations of assumptions, the credibility of the assumptions is improved. However, practical challenges to sensitivity analysis exist, such as the difficulty of interpreting sensitivity parameters, selecting appropriate models to produce unobserved confounders, and balancing identification strength with statistical power. These challenges highlight the need for careful implementation and transparency when using sensitivity analysis to support causal claims.

Consistency tests serve as an additional layer of scrutiny by evaluating whether an identification strategy produces results that align with theoretical expectations. Unlike sensitivity analysis, consistency tests

assess whether causal estimates behave logically in scenarios where the expected effect is known or can reasonably be inferred. Placebo tests, pre-trend analysis, and subgroup analysis are common forms of consistency tests that validate the credibility of causal assumptions. Although they do not directly prove the validity of identification strategies, they help detect inconsistencies that may indicate bias or misspecification of the model. The power of consistency tests depends on how strongly the expected outcomes should hold under the given assumptions. Therefore, it is important for researchers to select robust and theoretically justified testing conditions.

In practice, the results of assessing assumptions may not always reinforce our preferred conclusions. There are instances where sensitivity analyses and consistency tests reveal the implausibility of assumptions, raising doubts about the strength of causal claims. However, empirical analysis may still have value in expanding and deepening our knowledge, even if it does not provide definitive proof. In such cases, we apply the *law of decreasing credibility* (Proposition 3.3). Following its logic, we can relax the assumptions instead of abandoning causal inference altogether. We sacrifice the strength of conclusions for the sake of improved credibility. In this way, we may still learn meaningful insights. The implications of this principle will be discussed further in Chapter 6, where we explore strategies to adapt empirical conclusions in light of violations of assumptions.

6

Conclusions

Assumptions are the basis of causal inference, since they determine whether causal conclusions derived from empirical research are credible (Proposition 1.1) [1]. Unlike purely statistical analysis, which relies on associations between observed variables, causal inference seeks to establish cause-and-effect relationships. However, causal relationships cannot be directly observed (Propositions 2.1 and 3.4), which means that researchers must rely on assumptions to justify their claims (Proposition 3.1). These assumptions provide the logical structure necessary for causal identification (Proposition 3.2), ensuring that the inferences drawn from the data are not misleading or spurious. Therefore, the credibility of causal inference is based on a clear understanding of the assumptions underlying each method and research design.

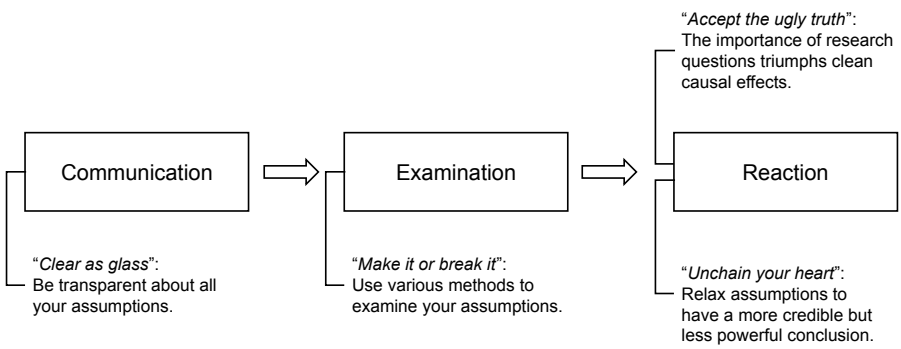
However, assumptions in causal inference present several key challenges. First, in causal identification, assumptions are often sufficient but not unique, which means that different sets of assumptions can lead to the same identification result (Proposition 4.2). This leads to the challenge of determining which set of assumptions is the most appropriate and credible in a given context. Second, researchers face a

[1] Please see the overview of all propositions in the monograph in Appendix C.

fundamental trade-off between the strength of their assumptions and the power of their conclusions (Proposition 3.3). If assumptions are too weak, identification may fail, but if they are too strong, conclusions may rely on unrealistic premises. Third, assumptions about potential outcomes, such as ignorability, are inherently untestable (Proposition 4.3). Furthermore, causal identification often depends on a set of interdependent assumptions (Proposition 4.4), where the plausibility of one assumption can hinge on the validity of others, complicating their examination. Given these challenges, researchers need a systematic framework for dealing with assumptions to enhance the credibility of causal inference.

In this monograph, I propose a systematic framework as depicted in Figure 6.1. The framework rests on open science spirit and the trade-off between strong assumptions and powerful conclusions. Researchers can follow the framework to clearly communicate, carefully examine, and prudently relax assumptions to enhance the credibility of their causal inference practices.

Figure 6.1: A Framework to Work with Assumptions in Causal Inference



The framework for dealing with assumptions in causal inference consists of three key stages: communication, examination, and reaction. The first stage, **communication**, emphasizes transparency about assumptions ("clear as glass"). Researchers should openly state and document key assumptions underlying their causal claims. A deep understanding of assumptions, as discussed in the monograph, is crucial for effective

communication. It helps researchers articulate what assumptions are being made, why they are necessary, and how they impact causal conclusions. Furthermore, this commitment to transparency aligns with recent calls in open science that advocate greater openness in research processes (Fecher and Friesike, 2014).

The second stage, **examination**, follows a "make it or break it" approach, where researchers apply various methods, such as sensitivity analysis and consistency tests, to examine the credibility of their assumptions. This stage helps to determine whether the assumptions hold under empirical testing or do not meet the credibility standards. Finally, the **reaction** stage involves a judgment call on whether to maintain or adjust the assumptions. Sometimes, researchers must recognize that the importance of answering a meaningful research question outweighs the pursuit of perfectly clean causal effects. In this case, the researcher may need to acknowledge that credible causal inference is unattainable ("accept the ugly truth"). Alternatively, they can "unchain their heart" by relaxing assumptions to favor more credible, but potentially weaker, conclusions. Together, these three stages provide a structured approach to handling assumptions, ultimately enhancing the credibility and transparency of causal inference.

Finally, I will discuss the relaxation of assumptions in causal inference, a practice that remains uncommon in empirical work. Relaxing assumptions inevitably has consequences, as reflected in the law of decreasing credibility (Proposition 3.3). Although weaker assumptions improve credibility, they also lead to less powerful conclusions. A compelling example is Manski *et al.* (1992), which examines the relationship between family structure during adolescence and the likelihood of high school graduation. The study highlights how different assumptions about this relationship yield different conclusions.

To address selection bias and prior information constraints, the authors employ three distinct approaches: nonparametric estimation under exogeneity, parametric latent-variable models, and the nonparametric bounds approach, each relying on progressively weaker assumptions. *Nonparametric estimation under exogeneity* assumes that family structure is independent of unobserved factors that affect graduation, allowing for precise estimates but depending on a strong and potentially

unrealistic assumption. *Parametric latent-variable models* incorporate strong prior information, imposing functional forms and assumptions of statistical independence to estimate a specific causal effect. In contrast, the *nonparametric bounds approach* makes no prior assumptions, resulting in wider but more credible bounds on the possible effects of family structure.

This variation in conclusions illustrates the law of decreasing credibility: the most precise estimates come from parametric models, but these rely on the strongest and least credible assumptions. Conversely, nonparametric bounds provide the weakest but most credible conclusions, as they impose weaker assumptions. This trade-off underscores the fundamental challenge in causal inference of balancing precision with credibility and highlights the importance of transparent communication of assumptions. In general, this example demonstrates how researchers can relax unrealistic assumptions while still obtaining meaningful conclusions.

Acknowledgements

The author thanks...

Appendices

A

Three Rules of Do-calculus

The three main rules of do-calculus are as follows:

- **Rule 1**: Remove incoming arrows to the intervened variable when it is independent of other variables, given some set of conditions.
- **Rule 2**: If there are multiple interventions, one of them can be removed if the intervened variable is independent of the others, given certain conditions.
- **Rule 3**: You can remove interventions on certain variables if they do not influence the desired outcome, given appropriate conditioning on other variables.

These rules enable us to compute the effect of interventions in complex causal models, making it possible to estimate causal effects from observational data by blocking or unblocking certain paths in the graph.

Let u's apply the three rules of do-calculus to the causal motifs (chain, fork, and collider) and explore how interventions can be evaluated. In the chain motif $D \to X \to Y$, the variable D influences Y through the mediator X. To evaluate the effect of an intervention on D, we apply

Rule 1 of do-calculus. After intervening on D (i.e., applying $\texttt{do}(D)$), the incoming arrows to D are removed, meaning $P(Y|\texttt{do}(D))$ depends only on the direct causal pathway through X. Formally, we compute:

$$\begin{aligned} P(Y|\texttt{do}(D)) &= \sum_x P(Y|X=x)P(X=x|\texttt{do}(D)) \\ &= \sum_x P(Y|X=x)P(X=x|D) \end{aligned} \tag{A.1}$$

In conclusion, the causal effect of D on Y or $\mathrm{P(Y|do(D))}$ is translated to conditional probabilities, which are directly calculable from data on D, X, Y.

In the fork motif $D \leftarrow X \rightarrow Y$, X is a common cause of both D and Y. In this case, conditioning on X blocks the indirect association between D and Y. When we intervene on X, using Rule 1 of do-calculus, we remove all incoming arrows into X, eliminating the natural causes of X. This allows us to evaluate the direct effect of X on both D and Y.

Using Rule 1, the do-calculus expression becomes:

$$P(Y|\texttt{do}(X)) = P(Y|X).$$

Because X is the common cause, intervening on X directly affects both D and Y, but there is no direct causation between D and Y, only association.

In the collider motif $D \rightarrow X \leftarrow Y$, X is a common effect of both D and Y. In this case, D and Y are independent unless we condition on X or its descendants. Applying do-calculus to this motif, we see that intervening on either D or Y does not introduce a direct causal relationship between them.

If we apply the do-operation on X, Rule 3 of do-calculus tells us that we should not condition on X, as it is a collider. Conditioning on X creates a spurious correlation between D and Y, which would otherwise be independent. Therefore:

$$P(Y|\texttt{do}(X)) \neq P(Y|X),$$

since conditioning on X introduces bias.

B

Proof for the Conditioning Strategy

We are interested in deriving the causal effect $P(Y \mid \texttt{do}(D))$ by applying the backdoor criterion. The backdoor criterion allows us to adjust for a set of variables X' (which corresponds to S in the original equation) that block all backdoor paths from D to Y.

Step 1: Initial Backdoor Adjustment Formula

According to the backdoor adjustment, the causal effect can be written as:

$$P(Y \mid \texttt{do}(D)) = \sum_{X'} P(Y \mid D, X')P(X' \mid D).$$

This formula shows that by conditioning on X', we adjust for all backdoor paths and sum over all possible values of X'.

Step 2: Factorizing the Joint Distribution

Next, we factor the joint distribution by introducing X (which represents $Pa(D)$, the parents of D):

$$P(Y \mid \texttt{do}(D)) = \sum_{X'} \sum_{X} P(Y \mid D, X', X)P(X' \mid D, X)P(X \mid D).$$

Step 3: Simplifying Using Conditional Independence

By the properties of the backdoor criterion, X' and D are conditionally independent given X. Therefore, we can simplify:

$$P(X' \mid D, X) = P(X' \mid X).$$

Substituting this back into the equation, we have:

$$P(Y \mid \texttt{do}(D)) = \sum_{X'} \sum_{X} P(Y \mid D, X', X) P(X' \mid X) P(X).$$

Step 4: Final Simplification

Since the backdoor criterion ensures that all confounding paths are blocked, we arrive at the final expression:

$$P(Y \mid \texttt{do}(D)) = \sum_{X'} P(Y \mid D, X') P(X'),$$

which matches equation (23.4). This demonstrates that the causal effect $P(Y \mid \texttt{do}(D))$ can be expressed using conditional probabilities based on the observational data.

C

Overview of Propositions

This appendix summarizes the propositions in the order in which they appear in the monograph.

- Proposition 1.1: The Golden Formula of Empirical Analysis.
- Proposition 2.1: The Fundamental Problem of Causal Inference.
- Proposition 3.1: The Essential Role of Identification in Causal Inference.
- Proposition 3.2: Identification as An Axiomatization Process.
- Proposition 3.3: The Law of Decreasing Credibility.
- Proposition 3.4: The Core Problem of Causal Inference.
- Proposition 4.1: The Role of Assumptions in Causal Inference.
- Proposition 4.2: Assumptions in Causal Inference are Sufficient.
- Proposition 4.3: Key Assumptions in Causal Inference Are Untestable.
- Proposition 4.4: Duhem-Quine Thesis in Causal Inference (assumptions are interdependent).

- Proposition 5.1: The Basic Logic of Sensitivity Analysis.
- Proposition 5.2: Determinants of the Power of Consistency Tests.

References

Abadie, A. (2021). "Using synthetic controls: Feasibility, data requirements, and methodological aspects". *Journal of Economic Literature*. 59(2): 391–425.

Abadie, A., A. Diamond, and J. 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, A. and G. W. Imbens. (2006). "Large sample properties of matching estimators for average treatment effects". *Econometrica*. 74(1): 235–267.

Andrews, I., J. H. Stock, and L. Sun. (2019). "Weak instruments in instrumental variables regression: Theory and practice". *Annual Review of Economics*. 11(1): 727–753.

Angrist, J. D. (2022). "Empirical strategies in economics: Illuminating the path from cause to effect". *Econometrica*. 90(6): 2509–2539.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. (2021). "Synthetic difference-in-differences". *American Economic Review*. 111(12): 4088–4118.

Athey, S., J. Tibshirani, and S. Wager. (2019). "Generalized random forests". *The Annals of Statistics*. 47(2): 1148.

Azoulay, P., C. Fons-Rosen, and J. S. G. Zivin. (2019). "Does science advance one funeral at a time?" *American Economic Review*. 109(8): 2889–2920.

Bai, J. (2009). "Panel data models with interactive fixed effects". *Econometrica*. 77(4): 1229–1279.

Bawa, K. and R. W. Shoemaker. (1987). "The effects of a direct mail coupon on brand choice behavior". *Journal of Marketing Research*. 24(4): 370–376.

Blackwell, M. (2014). "A selection bias approach to sensitivity analysis for causal effects". *Political Analysis*. 22(2): 169–182.

Bolton, R. N., P. K. Kannan, and M. D. Bramlett. (2000). "Implications of loyalty program membership and service experiences for customer retention and value". *Journal of the academy of marketing science*. 28(1): 95–108.

Bottmer, L., G. W. Imbens, J. Spiess, and M. Warnick. (2024). "A design-based perspective on synthetic control methods". *Journal of Business & Economic Statistics*. 42(2): 762–773.

Botvinik-Nezer, R., F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, *et al.* (2020). "Variability in the analysis of a single neuroimaging dataset by many teams". *Nature*. 582(7810): 84–88.

Breznau, N., E. M. Rinke, A. Wuttke, H. H. Nguyen, M. Adem, J. Adriaans, A. Alvarez-Benjumea, H. K. Andersen, D. Auer, F. Azevedo, *et al.* (2022). "Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty". *Proceedings of the National Academy of Sciences*. 119(44): e2203150119.

Carnap, R. (1936). "Testability and Meaning". *Philosophy of Science*. 3(4): 419–471.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press.

Cattaneo, M. D. and R. Titiunik. (2022). "Regression Discontinuity Designs". en. *Annual Review of Economics*. 14(1): 821–851. ISSN: 1941-1383, 1941-1391. DOI: 10.1146/annurev-economics-051520-021409. URL: https://www.annualreviews.org/doi/10.1146/annurev-economics-051520-021409 (accessed on 03/14/2024).

Chang, H. H., A. Mukherjee, and A. Chattopadhyay. (2023). "More voices persuade: The attentional benefits of voice numerosity". *Journal of Marketing Research*. 60(4): 687–706.

Chesher, A. and A. M. Rosen. (2017). "Generalized instrumental variable models". *Econometrica*. 85(3): 959–989.

Ching, A. T. and M. Osborne. (2020). "Identification and estimation of forward-looking behavior: The case of consumer stockpiling". *Marketing Science*. 39(4): 707–726.

Chintagunta, P. K. (2001). "Endogeneity and heterogeneity in a probit demand model: Estimation using aggregate data". *Marketing Science*. 20(4): 442–456.

Cohen, L. and C. J. Malloy. (2014). "Friends in high places". *American Economic Journal: Economic Policy*. 6(3): 63–91.

Davidson, R. and J. MacKinnon. (1993). "Estimation and Inference in Econometrics". *Tech. rep.* Oxford University Press.

De Haan, E., P. Kannan, P. C. Verhoef, and T. Wiesel. (2018). "Device switching in online purchasing: Examining the strategic contingencies". *Journal of Marketing*. 82(5): 1–19.

Dubé, J.-P., G. J. Hitsch, and P. E. Rossi. (2018). "Income and wealth effects on private-label demand: Evidence from the great recession". *Marketing Science*. 37(1): 22–53.

Duhem, P. and W. T. Scott. (1954). "The Aim and Structure of Physical Theory". *American Journal of Physics*. 22(7): 503–503.

Eggers, A. C., G. Tuñón, and A. Dafoe. (2024). "Placebo tests for causal inference". *American Journal of Political Science*. 68(3): 1106–1121.

Eggert, A., L. Steinhoff, and C. Witte. (2019). "Gift Purchases as Catalysts for Strengthening Customer–Brand Relationships". *Journal of Marketing*. 83(5): 115–132.

Eskin, G. J. and P. H. Baron. (1977). "Effects of price and advertising in test-market experiments". *Journal of Marketing Research*. 14(4): 499–508.

Fecher, B. and S. Friesike. (2014). *Open science: one term, five schools of thought*. Springer International Publishing.

Ferri-García, R. and M. d. M. Rueda. (2020). "Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys". *PloS one*. 15(4): e0231500.

Freyaldenhoven, S., C. Hansen, and J. M. Shapiro. (2019). "Pre-event trends in the panel event-study design". *American Economic Review*. 109(9): 3307–3338.

Giere, R. (1999). *Science without Laws*. Vol. 285. University of Chicago Press.

Gould, E., H. S. Fraser, T. H. Parker, S. Nakagawa, S. C. Griffith, P. A. Vesk, F. Fidler, D. G. Hamilton, R. N. Abbey-Lee, J. K. Abbott, *et al.* (2023). "Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology". *EcoEvoRxiv*.

Gruber, J. and B. Köszegi. (2001). "Is addiction "rational"? Theory and evidence". *The Quarterly Journal of Economics*. 116(4): 1261–1303.

Guo, T., S. Sriram, and P. Manchanda. (2021). "The effect of information disclosure on industry payments to physicians". *Journal of Marketing Research*. 58(1): 115–140.

Haavelmo, T. (1943). "The statistical implications of a system of simultaneous equations". *Econometrica*. 11(1): 1.

Hall, N. S. (2007). "RA Fisher and his advocacy of randomization". *Journal of the History of Biology*. 40(2): 295.

Hassan, M., J. Prabhu, R. Chandy, and O. Narasimhan. (2023). "When bulldozers loom: Informal property rights and marketing practice innovation among emerging market microentrepreneurs". *Marketing Science*. 42(1): 137–165.

Heckman, J. and B. E. Honore. (1990). "The Empirical Content of the Roy Model". *Econometrica*. 58(5): 1121–49.

Helmer, R. M. and J. K. Johansson. (1977). "An exposition of the Box-Jenkins transfer function analysis with an application to the advertising-sales relationship". *Journal of Marketing Research*. 14(2): 227–239.

Hempel, C. G. and P. Oppenheim. (1948). "Studies in the Logic of Explanation". *Philosophy of Science*. 15(2): 135–175.

Holland, P. W. (1986). "Statistics and causal inference". *Journal of the American statistical Association*. 81(396): 945–960.

Huang, Q., V. R. Nijs, K. Hansen, and E. T. Anderson. (2012). "Wal-Mart's impact on supplier profits". *Journal of Marketing Research*. 49(2): 131–143.

Hui, S. K., J. J. Inman, Y. Huang, and J. Suher. (2013). "The effect of in-store travel distance on unplanned spending: Applications to mobile promotion strategies". *Journal of Marketing*. 77(2): 1–16.

Huntington-Klein, N., A. Arenas, E. Beam, M. Bertoni, J. R. Bloem, P. Burli, N. Chen, P. Grieco, G. Ekpe, T. Pugatch, *et al.* (2021). "The influence of hidden researcher decisions in applied microeconomics". *Economic Inquiry*. 59(3): 944–960.

Imai, K., L. Keele, and T. Yamamoto. (2010). "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects". *Statistical Science*. 25(1): 51–71.

Imbens, G. W. and J. D. Angrist. (1994). "Identification and estimation of local average treatment effects". *Econometrica*. 62(2): 467–475.

Imbens, G. W. and T. Lemieux. (2008). "Regression discontinuity designs: A guide to practice". *Journal of econometrics*. 142(2): 615–635.

Imbens, G. W. and D. B. Rubin. (2015). "Causal Inference for Statistics, Social, and Biomedical Sciences". *Tech. rep.* Cambridge University Press.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Kim, S. J., R. J.-H. Wang, and E. C. Malthouse. (2015). "The effects of adopting and using a brand's mobile application on customers' subsequent purchase behavior". *Journal of Interactive Marketing*. 31(1): 28–41.

Koopmans, T. (1953). "The estimation of simultaneous linear economic relationships". *Studies in Econometric Method, Cowles Commission Monograph*. 14: 112–199.

Kuhn, T. S. (1997). *The Structure of Scientific Revolutions*. Vol. 962. University of Chicago Press, Chicago.

Lechner, M. *et al.* (2011). "The estimation of causal effects by difference-in-difference methods". *Foundations and Trends® in Econometrics*. 4(3): 165–224.

Lee, D. S. and T. Lemieux. (2010). "Regression Discontinuity Designs in Economics". en. *Journal of Economic Literature*. 48(2): 281–355. ISSN: 0022-0515. DOI: 10.1257/jel.48.2.281. URL: https://pubs.aeaweb.org/doi/10.1257/jel.48.2.281 (accessed on 03/14/2024).

Lewbel, A. (2019). "The identification zoo: Meanings of identification in econometrics". *Journal of Economic Literature*. 57(4): 835–903.

Li, X. and A. T. Ching. (2023). "Goodbye My Friends and Goodbye My Career: Evidence from the Movie Industry". *Available at SSRN 4575359*.

Lilien, G. L. (1979). "Exceptional paper—ADVISOR 2: Modeling the marketing mix decision for industrial products". *Management Science*. 25(2): 191–204.

Little, J. D. (1975). "BRANDAID: A marketing-mix model, part 1: Structure". *Operations Research*. 23(4): 628–655.

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Longoni, C., A. Bonezzi, and C. K. Morewedge. (2019). "Resistance to medical artificial intelligence". *Journal of Consumer Research*. 46(4): 629–650.

MacKinlay, A. C. (1997). "Event studies in economics and finance". *Journal of economic literature*. 35(1): 13–39.

Malani, A. and J. Reif. (2015). "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform". *Journal of Public Economics*. 124(C): 1–17.

Manski, C. (1997). "Monotone Treatment Response". *Econometrica*. 65(6): 1311–1334.

Manski, C. F. (1993). "Identification problems in the social sciences". *Sociological methodology*: 1–56.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F. (2013). *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press.

Manski, C. F. and J. V. Pepper. (2000). "Monotone Instrumental Variables: With an Application to the Returns to Schooling". *Econometrica*. 68(4): 997–1010.

Manski, C. F., G. D. Sandefur, S. McLanahan, and D. Powers. (1992). "Alternative estimates of the effect of family structure during adolescence on high school graduation". *Journal of the American Statistical Association*. 87(417): 25–37.

Matzkin, R. L. (2007). "Nonparametric identification". In: *Handbook of Econometrics*. Ed. by J. J. Heckman and E. E. Leamer. Vol. 6B. Elsevier. Chap. 73. 5307–5368. DOI: 10.1016/S1573-4412(07)06073-4. URL: https://www.sciencedirect.com/science/article/pii/S1573441207060734.

McCrary, J. (2008). "Manipulation of the running variable in the regression discontinuity design: A density test". *Journal of Econometrics*. 142(2): 698–714.

Miller, D. L. (2023). "An introductory guide to event study models". *Journal of Economic Perspectives*. 37(2): 203–230.

Moshary, S., B. T. Shapiro, and J. Song. (2021). "How and when to use the political cycle to identify advertising effects". *Marketing Science*. 40(2): 283–304.

Narang, U. and V. Shankar. (2019). "Mobile app introduction and online and offline purchases and product returns". *Marketing Science*. 38(5): 756–772.

Narayanan, S. and K. Kalyanam. (2015). "Position effects in search advertising and their moderators: A regression discontinuity approach". *Marketing Science*. 34(3): 388–407.

Oster, E. (2019). "Unobservable selection and coefficient stability: Theory and evidence". *Journal of Business & Economic Statistics*. 37(2): 187–204.

Palmatier, R. W. and A. T. Crecelius. (2019). "The "first principles" of marketing strategy". *AMS Review*. 9(1): 5–26.

Papies, D., P. Ebbes, and E. M. Feit. (2023). "Endogeneity and Causal Inference in Marketing". *World Scientific Book Chapters*: 253–300.

Park, N. K. (2004). "A guide to using event study methods in multi-country settings". *Strategic Management Journal*. 25(7): 655–668.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd. Cambridge University Press.

Pearl, J. and D. Mackenzie. (2018). *The Book of Why: The New Science of Cause and Effect*. Hachette UK.

Pham, M. T. (2013). "The seven sins of consumer psychology".

Popper, K. R. and G. Weiss. (1959). "The Logic of Scientific Discovery". *Physics Today*. 12(11): 53–54.

Quine, W. v. O. (1951). "Two dogmas of empiricism". In: *Can Theories be Refuted? Essays on the Duhem-Quine Thesis*. Springer. 41–64.

Rambachan, A. and J. Roth. (2023). "A more credible approach to parallel trends". *Review of Economic Studies*. 90(5): 2555–2591.

Rao, V. R. (1972). "Alternative econometric models of sales-advertising relationships". *Journal of Marketing Research*. 9(2): 177–181.

Robins, J. M., A. Rotnitzky, and L. P. Zhao. (1994). "Estimation of regression coefficients when some regressors are not always observed". *Journal of the American statistical Association*. 89(427): 846–866.

Rosenbaum, P. R. and D. B. Rubin. (1983a). "The central role of the propensity score in observational studies for causal effects". *Biometrika*. 70(1): 41–55.

Rosenbaum, P. and D. Rubin. (1983b). "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome". *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 45(2): 212–218.

Rosenbaum, P. and D. Rubin. (1985). "The bias due to incomplete matching". *Biometrics*. 41(1): 103–116.

Roth, J. (2022). "Pretest with caution: Event-study estimates after testing for parallel trends". *American Economic Review: Insights*. 4(3): 305–322.

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe. (2023). "What's trending in difference-in-differences? A synthesis of the recent econometrics literature". *Journal of Econometrics*. 235(2): 2218–2244.

Rubin, D. B. (1973a). "Matching to Remove Bias in Observational Studies". *Biometrics*. 29(1): 159.

Rubin, D. B. (1973b). "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies". *Biometrics*. 29(1): 185.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*. 66(5): 688.

Rubin, D. B. (1978). "Bayesian Inference for Causal Effects: The Role of Randomization". *The Annals of Statistics*. 6(1): 34–58.

Rutz, O. J. and G. F. Watson. (2019). "Endogeneity and marketing strategy research: An overview". *Journal of the Academy of Marketing Science*. 47: 479–498.

Schultz, R. L. and D. R. Wittink. (1976). "The measurement of industry advertising effects". *Journal of Marketing Research*. 13(1): 71–75.

Sharp, B. and A. Sharp. (1997). "Loyalty programs and their impact on repeat-purchase loyalty patterns". *International journal of Research in Marketing*. 14(5): 473–486.

Shriver, S. K., H. S. Nair, and R. Hofstetter. (2013). "Social ties and user-generated content: Evidence from an online social network". *Management Science*. 59(6): 1425–1443.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, *et al.* (2018). "Many analysts, one data set: Making transparent how variations in analytic choices affect results". *Advances in Methods and Practices in Psychological Science*. 1(3): 337–356.

Sorescu, A., N. L. Warren, and L. Ertekin. (2017). "Event study methodology in the marketing literature: an overview". *Journal of the Academy of Marketing Science*. 45: 186–207.

Stock, J. H. and M. Yogo. (2002). "Testing for Weak Instruments in Linear IV Regression". *Working Paper* No. 284. National Bureau of Economic Research. DOI: 10.3386/t0284. URL: http://www.nber.org/papers/t0284.

Sudhir, K. and D. Talukdar. (2015). "The "Peter Pan syndrome" in emerging markets: The productivity-transparency trade-off in IT adoption". *Marketing Science*. 34(4): 500–521.

Sudhir, K. (2001). "Competitive pricing behavior in the auto market: A structural analysis". *Marketing Science*. 20(1): 42–60.

Tamer, E. (2003). "Incomplete simultaneous discrete response model with multiple equilibria". *The Review of Economic Studies*. 70(1): 147–165.

Tamer, E. (2010). "Partial Identification in Econometrics". *Annual Review of Economics*. 2(1): 167–195.

Tan, Y.-C., S. R. Chandukala, and S. K. Reddy. (2022). "Augmented reality in retail and its impact on sales". *Journal of Marketing*. 86(1): 48–66.

VanderWeele, T. J. and P. Ding. (2017). "Sensitivity analysis in observational research: introducing the E-value". *Annals of internal medicine*. 167(4): 268–274.

Villas-Boas, J. M. (2004). "Consumer learning, brand loyalty, and competition". *Marketing Science*. 23(1): 134–145.

Von Neumann, J. and O. Morgenstern. (2007). "Theory of games and economic behavior: 60th anniversary commemorative edition". In: *Theory of games and economic behavior*. Princeton university press.

Wager, S. and S. Athey. (2018). "Estimation and inference of heterogeneous treatment effects using random forests". *Journal of the American Statistical Association*. 113(523): 1228–1242.

Wang, K. and A. Goldfarb. (2017). "Can offline stores drive online sales?" *Journal of Marketing Research*. 54(5): 706–719.

Wang, Y., M. Lewis, C. Cryder, and J. Sprigg. (2016). "Enduring effects of goal achievement and failure within customer loyalty programs: A large-scale field experiment". *Marketing Science*. 35(4): 565–575.

Xu, Y. (2017). "Generalized synthetic control method: Causal inference with interactive fixed effects models". *Political Analysis*. 25(1): 57–76.

Zhong, Z. (2022). "Chasing diamonds and crowns: Consumer limited attention and seller response". *Management Science*. 68(6): 4380–4397.