Causal Mediation Analysis

Geoffrey T. Wodtke
University of Chicago

Xiang Zhou
Harvard University

26 November 2024

DRAFT - DO NOT CIRCULATE

To our families, friends, and teachers

Contents

List of Figures

List of Tables

Acknowledgments

Writing this book has been both deeply rewarding and, on occasion, a true test of mental endurance. We are fortunate to have had the support of many individuals along the way. First, we would like to thank our families for their unwavering support and encouragement throughout this process. To Liz, Sydney, and Hailey, thank you for your patience with Geoff's many late nights reading, writing, editing, and coding. To Ken, Patti, Kat, Alec, and David, thank you for always believing in Geoff, despite his youthful mischief. To Qiang, Mason, and Charlie, thank you for your gracious support and understanding while Xiang spent numerous weekends working on this book.

We are also indebted to Jesse Zhou, Azim Shivji, and Daniel Klein for their assistance with coding and for engaging in numerous helpful discussions that improved the technical aspects of this project. Renato de Angelis provided essential support with reference management, for which we are very grateful. In addition to the help from Jesse, Azim, Daniel, and Renato, we also used chatGPT for assistance with coding and copy editing. Responsibility for all content and any potential errors in this book rests solely with us, however.

We deeply appreciate the insightful comments and feedback offered by Guanglei Hong, Felix Elwert, Chris Felton, and Fabian Kratz, all of whom helped us refine earlier drafts of the manuscript. A special thanks goes to Felix Elwert, whose thoughtful discussions throughout the writing process enriched several key sections of the book. We would also like to extend our gratitude to all the students who enrolled in SOCI 40258, Causal Mediation Analysis, at the University of Chicago. Their questions, feedback, and deep engagement in the class helped clarify and strengthen many of the concepts and examples presented in these pages.

Another special thank you is due to the staff and directors at the Stone Center for Research on Wealth Inequality and Mobility, whose support and encouragement in all of Geoff's academic endeavors made a significant difference to this project. We also want to express our gratitude to our former professors and advisers, especially those formerly affiliated with the University of Michigan's Quantitative Methodology Program. They taught both of us to think carefully and critically about statistical methods in the social sciences, skills that have been indispensable throughout our academic journey and to this project in particular.

Finally, Geoff would like to thank his neighbors in Chicago, who showed more interest in causal mediation than most folks outside the ivory tower, and whose enthusiasm and good cheer sustained him through this long and winding journey.

Chapter 1

Introduction

Causal mediation refers to a process in which one variable influences another through its impact on intermediate variables. This involves a sequence where a first variable affects a second, which then affects a third, and so on, forming a causal chain that links an initiating cause to a terminal outcome. For example, attending college may reduce the risk of unemployment, which leads to higher earnings, which may then improve mental health in adulthood. Participating in a job-training workshop can enhance practical skills and boost self-confidence, helping individuals to secure new employment. Consuming media that highlights the economic costs of immigration can trigger fear and anxiety, which may then erode support for certain immigration policies.

All these examples illustrate a process whereby an initial cause impacts intermediate variables that, in turn, influence an outcome. The initiating cause is often referred to as an exposure, treatment, or independent variable, while the intermediaries that transmit the effect of this cause to the outcome are known as mediators. The impact of the initial cause on the outcome that operates through its influence on a mediator or set of mediators is commonly termed an indirect effect. Conversely, the impact of the initiating cause on the outcome that bypasses the mediators of interest is called a direct effect (Hayes 2017; MacKinnon 2008; VanderWeele 2015).

Causal mediation analysis encompasses methods that evaluate direct and indirect effects and that delineate the intermediate mechanisms connecting a cause to an outcome. It seeks to uncover not just *whether* but also *why* a putative cause influences an outcome by quantifying the processes through which an effect of interest operates. While many scientific studies aim only to establish the existence, direction, or magnitude of a causal effect from one variable to another, mediation analysis transcends this narrow focus: it additionally probes the underlying mechanisms in an effort to illuminate exactly how causes produce their effects.

This book provides a comprehensive introduction to causal mediation analysis, tailored for applications in the social sciences. It uses precise notation and clear visual diagrams to guide readers from basic definitions of effects, via minimally necessary assumptions, to cutting-edge estimation procedures. With an emphasis on both theoretical foundations and practical implementation, the text blends technical material with many empirical illustrations, all implemented using widely available data and software. Our goal is to furnish readers with the knowledge and skills needed to conduct a defensible analysis of causal mediation across a wide range of scenarios. This includes the ability to define, identify, and estimate a variety of different direct and indirect effects, interpret these effects accurately, and critically assess the assumptions required for valid inference. Throughout the book, we highlight both the promise and the pitfalls of causal mediation analysis, carefully outlining what can–and cannot–be inferred from empirical data, under what assumptions, and with what level of certainty.

Although the concept of mediation is straightforward—intermediate variables transmit the effect of an initial cause to an outcome—investigating this process empirically can be rather complex. This book serves as an introductory guide, but it does presume some knowledge of graduate-level statistics and research methods, as well as familiarity with regression and general linear models (Fox 2015; Wooldridge 2020). Prior exposure to the foundations of causal inference, including the counterfactual framework and graphical causal models (Angrist and Pischke 2009; Hernan and Robins 2020; Morgan and Winship 2014), would be helpful but not essential. We introduce these topics early on and revisit them throughout the book. To make the material as accessible as possible, technical definitions of important concepts, models, and assumptions are paired with intuitive explanations and practical examples wherever possible.

1.1 Mediation in the Social Sciences

Causal mediation analysis is central to research in the social sciences, including sociology, psychology, political science, public health, criminology, economics, and public policy evaluation. Across these disciplines, examples of mediation abound: environmental stimuli prompt learning that shapes behavior; family background influences educational attainment, which subsequently affects earnings in adulthood; marriage facilitates greater investment in children, which improves their performance at school; microbial exposures result in infections that lead to death; tobacco taxes decrease smoking, which then reduces cancer incidence; and so forth. This section introduces several areas of social scientific inquiry involving causal mediation, topics that we will revisit and elaborate on throughout the book.

1.1.1 The Effects of Education on Mental Health

The mental health of a population serves as a "social mirror," reflecting the cumulative human consequences of different environmental contexts, economic conditions, and personal histories (Avison et al. 2007). On an individual level, mental health problems can disrupt work and relationships, thereby exacerbating isolation and further reducing life satisfaction. These challenges also impose significant economic costs, which are evidenced by lower productivity, higher healthcare spending, and strained public resources.

In the fields of sociology, economics, and public health, education is widely thought to improve mental health (Heckman et al. 2018; Hout 2012; Mirowsky 2003). For example, numerous studies document a negative correlation between higher education and depression, with strong indications that this relationship is causal (Adams et al. 2003; Warren 2009; Yan and Williams 1999). A more advanced education therefore appears to reduce depression, but important questions remain. In particular, through what mechanisms does education exert its beneficial effects on mental health? What processes explain the lower levels of depression that follow from higher levels of educational attainment?

Education may improve mental health through a number of different pathways. For example, higher education might prevent depression by insulating individuals from the economic and psychological stressors associated with unemployment. Similarly, adults with higher levels of education often experience greater demand for their skills in the labor market, leading to higher-paying jobs compared to their counterparts with lower levels of education (Card 1999; Hout 2012). The greater incomes that accrue to the highly educated may subsequently improve mental health by helping to fulfill their basic needs and providing better access to healthcare, while also boosting self-esteem and enhancing personal agency (Mirowsky and Ross 1990; Silverstone and Salsali 2003). The influence of education on mental health might even extend beyond its impact on employment and earnings. An advanced education could also reduce depression by providing access to health-related information, promoting healthier lifestyles, or enlarging support networks (Lee 2011; Mirowsky 2003).

Does employment stability transmit the effect of education on depression? How important is the explanatory role of higher income? Can we differentiate the impact of these economic processes from the influence of other factors, such as health information, lifestyle choices, and social support? These questions, which probe the mechanisms through which one variable affects another, are the realm of causal mediation analysis.

1.1.2 Media Discourse, Public Opinion, and Immigration

Throughout history, attitudes toward immigration have significantly influenced politics, governance, and inter-group conflict (Glazer and Moynihan 1970; Omi and Winant 2014; Portes and Rumbaut 2006). Serving often as a catalyst for political polarization, immigration has consistently shaped public opinion, national policy, and social cohesion during periods of heavy population movement. The recent intensification of these debates in North America and Western Europe, spurred by globalization and its concomitant demographic shifts, underscores the persistent contentiousness surrounding immigration.

In the fields of political science, sociology, and psychology, understanding the determinants of attitudes toward immigration is a central concern. These attitudes are highly volatile, with opposition to immigration waxing and waning from one decade to the next (Simon and Lynch 1999). Media discourse and messaging from political elites are widely recognized as key influences on public sentiments. In particular, news stories and political communications that emphasize the perceived negative impacts of immigration on jobs, crime, taxes, public services, or social cohesion tend to amplify public opposition (Brader et al. 2008).

But what psychological processes account for these effects? One perspective contends that media highlighting the harms of immigration may simply change individual beliefs about the severity of the issue (Scheve and Slaughter 2001; Sniderman et al. 2000). According to this view, exposure to negative media framing alters perceptions of the costs and benefits associated with immigration, thereby intensifying opposition. Alternatively, exposure to media emphasizing the harms of immigration could provoke fear, anxiety, or distrust, and this emotional response, particularly potent when the immigrants in question are members of a denigrated racial group, might fuel opposition to immigration (Brader et al. 2008; Marcus et al. 2005, 2000).

Does exposure to media that highlights the costs of immigration heighten opposition by triggering fear and anxiety? Or does the influence of negative framing stem from changes in beliefs about the costs and benefits of immigration? Disentangling the role of beliefs versus emotions in transmitting the effects of media framing on public opinion is yet another task for causal mediation analysis.

1.1.3 Workforce Development Programs and Employment

Unemployment is a critical social problem with far-reaching impacts on both individuals and communities. High unemployment rates are associated with increased poverty and greater mortality at the population level (Brady 2009; Roelfs et al. 2011). The personal toll of unemployment is also severe, as it can lead to psychological distress, diminished self-esteem, and frayed social relationships (Jahoda 1982; Paul and Moser 2009). Addressing unemployment and facilitating a return to work are therefore crucial not only for the economic vitality of society but also for the personal well-being of its members.

Re-employment programs can help mitigate the negative impacts of job loss, enhance productivity, and foster a sense of purpose and belonging among individuals who are out of work. Research in economics, public policy, and program evaluation has explored a variety of workforce development interventions designed to assist unemployed workers with re-entering the labor force (Barnow 1987; Bloom et al. 1997; Holzer 2013). One strategy is to enroll these individuals in community-based workshops that provide training in a variety of skills, such as job search techniques, resume construction, interviewing demeanor, and problem-solving. Additionally, these workshops may offer programming to help boost self-esteem and self-efficacy. Participation has been shown to facilitate reemployment and increase earnings, though the effects can sometimes be modest (Vinokur et al. 1995; Vinokur and Schul 1997).

How do these interventions succeed in increasing employment? Do they equip participants with tangible skills that employers value and help them to find new work, or do they primarily instill confidence, which sustains individuals through a prolonged job search and fosters resilience against setbacks? Could the workshops be modified to enhance their impact on employment by concentrating more on psychological well-being or practical skills? Causal mediation analysis can reveal how employment programs achieve their impacts and inform the development of more effective interventions in the future.

1.2 The Promise of Causal Mediation Analysis

Causal mediation analysis is central to research in the social sciences for several compelling reasons (Glynn 2021; MacKinnon et al. 2007; VanderWeele 2015). First, it enriches our understanding of social phenomena. For example, discerning why participation in a job training workshop leads to increased employment and earnings—whether through improved self-efficacy, the acquisition of tangible skills, or both—deepens our understanding of how the social world operates. Such knowledge not only satisfies a fundamental human curiosity but also supports informed decision-making and lays a foundation for future learning and discovery.

Second, causal mediation analysis is essential for evaluating and refining theory. Many theories in the social sciences posit causal chains linking different variables or events. Sometimes, competing models propose different mechanisms for the same effect, from an initial cause to a terminal outcome. By analyzing these hypothesized mechanisms, we can test, refine, and adjudicate between theories more rigorously. This enhances the reliability and applicability of our theoretical models, which can then support more accurate predictions and better decision-making outside the social science laboratory.

Third, causal mediation analysis can help improve interventions and public policies. For example, if evidence suggests that a job training workshop succeeds at increasing employment primarily by boosting participant self-efficacy, future iterations of the program could focus more on confidence-building exercises to potentially enhance its effectiveness. Conversely, mediation analysis might reveal components of the program that are ineffective or even detrimental, such as activities that inadvertently heighten anxiety or despair, which may then inhibit the search for a new job. These components could be discarded or marginalized in an effort to maximize the program's beneficial impact. In general, knowledge of the causal mechanisms through which an intervention affects an outcome can aid in optimizing its impact and in diagnosing failures to achieve intended results.

Relatedly, mediation analysis can also inform the development of new policy interventions by identifying more feasible targets for change. While addressing the root cause of a problem is ideal, it is often impractical. Consider the example of education and mental health. If research shows that education primarily improves mental health by increasing income, then directly boosting income through targeted financial assistance to workers or families might be a more practical solution for realizing the same benefits compared to encouraging further education, which can be logistically complicated. Mediation analysis can point toward more promising and actionable points of intervention when directly addressing the initial cause proves impractical or infeasible.

In addition, causal mediation analysis can contribute to understanding how the effects of interventions might vary across different populations, time periods, or settings. By identifying specific mechanisms that drive causal effects, researchers can better predict where and for whom these interventions are likely to be more versus less effective. This information can be useful when tailoring interventions to specific target populations or transporting them from one setting to another.

Finally, mediation analysis can bolster the credibility of inferences about the causal effect of one variable on another. In the social sciences, where conducting randomized experiments can be logistically challenging or unethical, researchers often depend on a variety of different study designs to investigate causation. By triangulating these different types of evidence, we can develop a more robust understanding of causal relationships between variables. Mediation analysis contributes to this pluralistic approach by tracing the entire process—from the initial cause, through intermediate variables, to the final outcome—thereby strengthening the support for causal inferences about any one link in the chain.

Despite the promise of causal mediation analysis for deepening understanding, evaluating theory, refining interventions, and bolstering the credibility of causal inferences, tracing causal processes remains challenging in practice. In most applications, the assumptions required to learn about causal mediation from empirical data are considerably stronger than those needed to infer relationships between only an initial cause and a terminal outcome (VanderWeele 2015). Furthermore, designing experiments to study mediation poses additional challenges. While randomized experiments are effective and widely used for assessing the impacts of one variable on another, their capacity to reveal how these effects are transmitted through intermediate factors is comparatively limited, as experiments specifically aimed at uncovering mediation are more complex and difficult to implement.

Nevertheless, understanding mediation is crucial for social science inquiry. Moreover, reasoning and learning about causal chains are central components of human intelligence in general. From as early as age three, individuals begin to interpret the world around them in terms of causal processes (MacKinnon 2008; Shultz 1982). Successfully navigating our social and natural environments is almost unimaginable without the ability to learn and apply knowledge of mediation—forecasting and anticipating the sequences of events triggered by an initial cause, its intermediate consequences, and ultimate outcomes. Although the scientific investigation of mediation is fraught with challenges and uncertainties, it remains indispensable to human understanding, decision-making, and action, not only in the social sciences but also in everyday life.

1.3 A Brief History

Over the past several decades, a "causal revolution" has transformed research in statistics and the social sciences, driven by seminal works from Donald Rubin (1974; 1983), James Robins (1986) and Judea Pearl (1995; 2009). Rubin sparked the revolution by introducing the counterfactual framework for conceptualizing causation and explicitly distinguishing it from statistical association. Robins then extended this framework to a wider range of applications, such as those involving longitudinal data.

The counterfactual framework established a rigorous notation for defining causal relationships in terms of "potential outcomes," which envision different, hypothetical states of the world to determine what would have occurred if a variable or set of variables had differed from their actual state. With this approach, if an outcome would have differed had some exposure been other than it was, that exposure is considered a cause of the outcome, though not necessarily the sole cause. The introduction of potential outcomes recast causal inference as a missing data problem. Since we can never observe the outcome that would have occurred had an exposure differed from its factual state, we are always missing a key piece of information needed to infer causality with certainty. Thus, the challenge of causal inference lies in trying to learn about outcomes that cannot be observed, using only the data that is observable.

Following the introduction of the counterfactual framework, Pearl accelerated the causal revolution by developing a graphical framework paired with a "causal calculus" for analyzing causation. This approach allows researchers to visually represent causal relationships, deduce potential outcomes from a broader causal system, and derive testable implications for observable data. These innovations also help to identify the specific conditions under which we can learn about counterfactual outcomes from the empirical data available, along with the necessary calculations to achieve this learning. Together, graphical causal models and Pearl's calculus have enhanced our ability to precisely determine when and how statistical associations in observed data imply causation–or when they do not.

Before the causal revolution, efforts to quantify mediation started with Wright's (1921; 1934) development of linear path analysis. This approach uses simple diagrams and linear models to estimate and visualize direct and indirect relationships among a set of variables. Linear path analysis gained traction in the social sciences during the 1960s and 1970s, where it was further developed by Duncan (1966), Goldberger (1972; 1973), and Blalock (1971). Later, Alwin and Hauser (1975) systematized procedures for decomposing a total effect into direct and indirect components using linear models, while Sobel (1982) developed procedures for drawing statistical inferences from these types of decompositions. In a highly influential article, Baron and Kenny (1986) introduced a closely related methodology, which involves estimating a series of linear models and applying statistical tests to certain coefficients in order to evaluate mediation. More recently, texts by Hayes (2017) and MacKinnon (2008) have widely disseminated these approaches across the social sciences, focusing largely on methods that stem from the tradition of analyzing mediation with linear models.

All of these approaches, from Wright's (1934) early work to Duncan (1966), and from Baron and Kenny (1986) to contemporary texts like Hayes (2017) and MacKinnon (2008), agree that mediation is fundamentally a causal concept. For example, Baron and Kenny (1986, pg. 1173) define mediation as "the generative mechanism through which the focal independent variable is able to influence the dependent variable of interest." Similarly, Alwin and Hauser (1975, pg. 39) describe indirect effects as "those parts of a variable's total effect which are transmitted...by variables specified as intervening between the cause and effect of interest." Likewise, MacKinnon (2008, pg. 8) explains that "in a mediation model, the independent variable causes the mediator which then causes the dependent variable," while Hayes (2017, pg. 7) describes mediation as occurring when "variation in [an exposure] X causes variation in one or more mediators M, which in turn causes variation in [an outcome] Y."

Despite the broad consensus that mediation inherently involves causation, most earlier works did not adopt a formal notation or conceptual framework that clearly distinguished between causation and mere association. Prior approaches to mediation analysis in the social sciences have typically defined and analyzed mediation within the confines of statistical models that blur the distinction between causal versus associational relationships, forgoing the innovations of the causal revolution described previously.

In recent years, however, methodological research on mediation analysis has fully embraced the counterfactual framework and graphical causal models (e.g., Pearl 2001; Robins and Greenland 1992; VanderWeele

2011b), leading to significant breakthroughs, important new insights, and a surge of novel methods. Unfortunately, much of this innovative material remains buried within specialist journals, often presented in highly technical language that may be inaccessible to applied researchers who could greatly benefit from these advances. Even in texts aimed at a more general audience that draw heavily on these recent advances, such as VanderWeele's (2015) outstanding treatise on "Explanation in Causal Inference," the authors sometimes still choose to forgo the modern tools of causal analysis, opting instead to "describe as many of the methods and assumptions as possible without requiring specific appeal to the notation of counterfactual-based logic" (p. xi).

In contrast, we consider the counterfactual framework and causal graphs as indispensable tools for mediation analysis. Without an explicit notation and causal calculus, our view is that any approach to analyzing mediation remains incomplete and is liable to generate confusion regarding the objectives of the analysis and the conditions needed to achieve them. Embracing the modern tools of causal analysis broadens the range of theoretical questions about mediation that can be precisely articulated, clarifies the role of empirical data versus unverifiable assumptions in addressing these questions, and unlocks new tools for extracting answers to them.

1.4 Our Distinctive Approach

Building on the innovations that propelled the causal revolution, our approach to mediation analysis comprises four basic steps:

1. **Define:** clearly articulate a causal estimand using potential outcomes, which should faithfully represent the substantive aims of the research and require no commitment to any particular statistical model.

2. **Identify:** connect the causal estimand, which involves counterfactuals that cannot be observed, to an empirical quantity that can be learned from observable data under a set of precisely stated assumptions.

3. **Estimate:** collect these data and use them to estimate the empirical quantity of interest, assessing any uncertainty that afflicts the estimation.

4. **Scrutinize:** critically evaluate the assumptions required for the data to yield insights about the causal estimand, which is essential for judging the credibility of the evidence provided.

Following the framework proposed by Lundberg et al. (2021), our first step involves defining a causal estimand, or a set of estimands, using potential outcomes and the counterfactual framework. An estimand is simply the object of scientific inquiry—it represents the theoretical quantity we would aim to determine if it were possible to observe all potential outcomes from each of the counterfactual worlds envisioned. This step translates the substantive goals of the research into precise objects for inference that exist independently of any statistical model and clearly distinguish causation from association.

The second step in our approach involves linking the causal estimand, defined in terms of quantities that cannot be observed, to an empirical estimand that consists solely of observable quantities. Essentially, this step connects the causal estimand of interest to a function of empirical data. Causal estimands defined in terms of counterfactuals can only be linked with empirical quantities through assumptions about how the unobservable data relate to the data we can observe. These assumptions must be articulated with precision and transparency, as the empirical data are informative about the causal estimand only when they hold true.

Although there are a variety of ways to express these assumptions, causal graphs are particularly effective and accessible, and we rely on them throughout the book.

The third step involves collecting data and selecting an estimation strategy to learn about the empirical quantity of interest. By decoupling the causal estimand and its empirical counterpart from the estimation process, researchers are afforded a range of options for learning from data, each with distinct strengths and weaknesses. Defining the target of scientific inquiry independently of a statistical model allows researchers the freedom to choose the most suitable method for their data, without being tethered to any single approach to estimation. Estimation can proceed without relying on any statistical model whatsoever, if the data permit, or it can involve a variety of models that offer different levels of flexibility in approximating the true but unknown distribution of the observed data. This approach enables researchers to conduct mediation analysis with almost any type of data, no matter the complexity of the relationships or the level of measurement among variables.

It also stands in stark contrast to conventional practices in the social sciences, where researchers analyzing mediation often leap straight to a particular statistical model without separately defining and identifying a causal estimand as the ultimate target of inference. Typically, the target of inference is defined solely within the confines of a particular model for the observed data–for example, as a product of coefficients from several linear models (Baron and Kenny 1986; Hayes 2017) or as a difference in re-scaled coefficients from a series of logit models (Karlson et al. 2012; MacKinnon 2008). However, this conventional approach is suboptimal for a number of reasons. In bypassing a model-independent estimand to focus solely on a single statistical model for the observed data, researchers obscure the true objectives of their analysis and the assumptions needed to achieve them, limit the exploration of alternative models that might offer better performance, restrict their ability to accommodate complex relationships and data structures, and hinder the accumulation of scientific knowledge across studies using different modeling strategies to examine the same phenomena (Lundberg et al. 2021).

The final step involves a critical examination of the assumptions needed to link the causal estimand to its empirical counterpart and the estimation strategy. Because it is often impossible or impractical to ensure all necessary assumptions are met by design, it is crucial to assess the sensitivity of causal inferences to potential violations of the assumptions on which they are based. In practice, this involves strategically altering the assumptions and observing the impact of these changes on the results, evaluating whether key conclusions of the analysis remain valid when the assumptions are modified. If minor changes to the assumptions result in significant alterations in the results, the original conclusions may not be very robust. Conversely, if the conclusions hold despite these modifications, researchers can have greater confidence in the reliability of their findings.

To summarize, our approach to causal mediation analysis unfolds in four key steps: define, identify, estimate, and scrutinize. It begins by precisely stating the causal estimand—the central quantity of interest—completely independent of any statistical model. This is followed by linking the causal estimand, which cannot be observed directly, to an empirical quantity based on a set of clearly stated assumptions. The next step involves estimating this empirical quantity from data and quantifying the uncertainty that afflicts the estimates. It concludes with a critical evaluation of assumptions to determine if the causal inferences are robust to their potential violation. This approach contrasts with conventional approaches to mediation analysis in the social sciences, which often conflate associational and causal notation, rely solely on model-based definitions of effects, obscure the assumptions required to learn about them from empirical data, and consequently, neglect to adequately scrutinize these assumptions.

The approach we adopt offers several advantages for studying mediation, compared to methods that are conventional at present in the social sciences. It sharply distinguishes association from causation, ensuring clarity and precision about the objectives of the research. It focuses on model-independent definitions and identification of target estimands, unlocking the use of a diverse array of competing estimation procedures. This flexibility allows researchers to accommodate many different types of relationships among variables and a variety of different data structures. Moreover, our approach ensures transparency regarding the assumptions underpinning inferences about causal mediation, and it facilitates assessment of how sensitive these inferences are to potential violations of the assumptions on which they are based. Together, these advantages form a solid foundation for generating more credible evidence of causal mediation and for accumulating scientific knowledge about causal processes.

1.5 An Outline of the Book

In the ensuing chapters, we provide a comprehensive introduction to causal mediation analysis, with applications to research in the social sciences. We begin by outlining fundamental concepts and relatively simple applications, and then gradually tackle more complex scenarios and the increasingly sophisticated methods they demand. Throughout these chapters, we focus on a running example that examines the impact of college attendance on depression at midlife, and whether this relationship is mediated by unemployment and/or income. This example is revisited in each chapter to ensure continuity and to highlight the subtle differences between methods for analyzing distinct types of mediation. To appeal to an interdisciplinary audience and enrich the narrative, each chapter also concludes with a unique, stand-alone empirical illustration. These additional examples are drawn from a wide range of disciplines in the social sciences.

Chapter 2 lays the foundation for the remainder of the book by introducing concepts and notation that are essential for analyzing causal mediation. We first outline the counterfactual framework, potential outcomes, and graphical causal models, which form the backbone of any causal analysis. Using these tools, we conceptually define the total effect of an exposure on an outcome and explain the "fundamental problem of causal inference" (Holland 1986). We then explain how potential outcomes and causal diagrams can be used to distinguish between a direct effect of the exposure that bypasses a mediator of interest, and an indirect effect that operates specifically through a causal chain involving the mediator. After conceptually defining direct and indirect effects, we introduce the "fundamental problem of causal mediation," a challenge that is related to the fundamental problem of causal inference but is even more difficult to resolve. This chapter serves as a conceptual primer that illustrates the key ideas, notation, and principles, as well as the challenges, that underlie analyses of causal mediation.

In Chapter 3, we introduce methods for analyzing causal mediation in applications involving a single mediator, where the direct and indirect effects of interest may be confounded by baseline covariates. Baseline confounding occurs when extraneous variables, which are either measured before the exposure or are otherwise unaffected by it, distort the apparent effects of the exposure and/or the mediator, leading to biased inferences. This chapter outlines the conditions necessary for accurately analyzing total, direct, and indirect effects in applications with a single mediator and baseline confounding, and it provides a comprehensive overview of different estimation procedures, including those based on linear models, simulation methods, and inverse probability weighting. To illustrate these methods, we analyze data from a study that assessed the impact of a job training workshop on employment (Vinokur et al. 1995; Vinokur and Schul 1997), focusing specifically on how psychological factors like self-efficacy may mediate these effects.

Chapter 4 addresses a more complex but still common scenario in the social sciences: causal mediation analysis focusing on a single mediator in the presence of exposure-induced confounding. This type of confounding occurs when variables affected by the exposure distort the apparent effects of the mediator on the outcome, which can also lead to biased inferences. In this chapter, we introduce an alternative approach for analyzing how a single mediator may account for the effect of an exposure on an outcome when exposure-induced confounders are present. We discuss modified regression, simulation, and weighting procedures as viable estimation strategies. And to illustrate these methods, we re-analyze cross-national data from a study that examined the economic origins of gender roles (Alesina et al. 2013). Specifically, we analyze whether historical cultivation practices (i.e., the use of plow agriculture) influenced contemporary patterns of female political participation directly, apart from their impact on economic development, all while accounting for a potentially important exposure-induced confounder–the degree to which a country's current government is authoritarian versus democratic.

Chapter 5 covers analyses of causal mediation with multiple mediators. In this chapter, we outline methods for decomposing the total effect of an exposure on an outcome into an indirect effect, which operates through a set of mediators, and a direct effect that circumvents them all. We also show how to decompose a total effect into a series of path-specific effects, which capture the unique explanatory role of each mediator in a causal sequence. To estimate these effects, we outline several strategies, including regression-imputation and weighting approaches. We demonstrate these methods by re-analyzing data from an experimental study of public opinion that examined how negative media framing affects opposition to immigration (Brader et al. 2008). Using these data, we attempt to untangle the role of emotions versus beliefs as potential mediating factors that connect discourse in the news to public opinion.

Chapter 6 explores the application of machine learning to enhance causal mediation analysis. Machine learning is a burgeoning field of statistics and computer science that involves training algorithms to inductively identify relationships and patterns in data. When used to analyze causal mediation, this approach combines carefully constructed estimating equations with these data-adaptive algorithms to produce robust estimates of direct and indirect effects. We initiate the discussion by addressing the problem of model misspecification, which often complicates analyses of causal mediation. We then introduce the concept of "multiple robustness" and explain the advantages of machine learning methods for causal inference. Lastly, we present a collection of multiply robust estimation procedures for use in analyses of causal mediation. To demonstrate these methods, we re-analyze data from a study that examined the inter-generational pathways through which exposure to political violence shapes descendants' political attitudes (Lupu and Peisakhin 2017). This chapter highlights the challenges of model misspecification, outlines the advantages of multiply robust approaches to estimation, and demonstrates the utility of machine learning for mediation analysis.

In Chapter 7, we explore the design and analysis of experimental and quasi-experimental studies specifically aimed at uncovering causal mediation. The chapter discusses a number of novel experimental designs, which are scarcely applied in the existing literature but hold potential for advancing research on mediation. These include joint, parallel, and multi-arm randomization designs, as well as experiments utilizing randomized encouragements and pre/post designs with repeated measures of the mediator.

Finally, Chapter 8 concludes by reflecting on limitations and omissions. It charts the frontier of the methodological literature on causal mediation and suggests potentially fruitful directions for the future, aiming to inspire ongoing dialogue and investigation in this vital area of social science research.

1.6 Statistical Software

The widespread adoption of causal mediation analysis is facilitated by the availability of software for implementation. In support of this, all the data and code used throughout this book, complete with custom functions and detailed documentation for implementing the methods we discuss and illustrate, are accessible at `https://github.com/causalMedAnalysis`. We provide code in both R and Stata, which are among the most popular statistical software packages in the social sciences. These resources are designed to equip readers with the tools necessary to replicate and explore the examples presented in this book. Additionally, they provide the means for readers to adapt and apply all the methods we cover to their own research. Moving forward, we are committed to continually updating and enhancing the capabilities and functionality of the code, ensuring its ongoing relevance and utility.

Chapter 2

Foundations of Causal Inference

It is commonly hypothesized that attending college affects mental health (Hout 2012; Heckman et al. 2018; Mirowsky 2003), contributing to lower levels of depression and greater life satisfaction (Warren 2009; Adams et al. 2003; Miech et al. 1999). But what, exactly, does it mean to assert that attending college causally affects mental health? How can we precisely define these effects and distinguish them from mere statistical associations? More importantly, how can we determine when observed associations reflect causal relationships and when they do not? These are some of the fundamental questions at the heart of any inquiry into causality.

In this chapter, we lay the foundation for understanding causal mediation analysis by introducing several key concepts and tools. We begin with the counterfactual framework (Rubin 1974; Rosenbaum and Rubin 1983; Holland 1986), a cornerstone of modern causal inference. This framework defines causal effects by comparing potential outcomes–the values an outcome would take for an individual under different, possibly counterfactual, states of the world. According to this approach, an exposure is considered a cause of an outcome if the outcome's potential values would differ under one level of exposure versus another, which may not align with the level of exposure an individual actually experiences in reality. Defining causal effects through comparisons of counterfactual scenarios highlights a central challenge in analyses of causality, known as "the fundamental problem of causal inference." This problem arises because we never observe the potential outcomes that would have occurred under a level of exposure that an individual did not actually experience. In other words, we never observe all the pieces of information that define a causal effect. The counterfactual framework, then, recasts causal inference as a problem of missing data, where the primary task is to overcome this problem by identifying counterfactual quantities using only the information that can be observed.

Next, we introduce directed acyclic graphs (DAGs; Elwert 2013; Pearl 1995, 2009), which allow researchers to visually represent causal relationships among variables. DAGs not only offer a clear and concise way to graphically depict these relationships, they also enable researchers to derive potential outcomes from a broader causal system, to identify the conditions under which causal effects—defined in terms of these potential outcomes—can be learned from the observable data, and to formulate testable implications. In this way, DAGs serve as a powerful tool for addressing the fundamental problem of causal inference. We illustrate how DAGs complement the potential outcomes framework and facilitate the analysis of more complex causal relationships, such as those encountered in causal mediation analysis.

The chapter then transitions to an introductory discussion of causal mediation, focusing on the conceptual distinction between direct versus indirect causation. We initiate this discussion by outlining a simple graphical model of mediation, where direct and indirect causation are distinguished by different paths in a

DAG. We then introduce several types of potential outcomes that involve not only counterfactual states of the exposure but also of a putative mediator. These more complex potential outcomes help formalize the concept of mediation and clarify the distinction between direct and indirect influences on an outcome. With these concepts established, we conclude by discussing "the fundamental problem of causal mediation," another challenge arising from missing data on certain potential outcomes. This problem is similar to the fundamental problem of causal inference mentioned previously, but even more challenging to resolve. Our discussion of this problem prefaces a brief preview of other challenges that commonly afflict analyses of causal mediation, such as the difficulty of designing experiments to investigate direct and indirect effects, the pervasive need to adjust for confounding variables, and the complexities of untangling multiple mediators–challenges that we will explore in detail throughout the remainder of the book.

To reinforce the abstract concepts and ideas discussed in this chapter, we illustrate them with an example based on the 1979 National Longitudinal Survey of Youth (NLSY; Bureau of Labor Statistics 2019), examining the effect of college attendance on depression at midlife. By the end of this chapter, readers will have a solid foundation in the formal language and logic of causal inference, preparing them to translate, define, and reason about causal arguments, statements, and hypotheses in social science research. This foundation sets the stage for more detailed explorations of causal mediation in the chapters that follow.

2.1 Measures of Association

The distinction between causation and association is fundamental to causal inference. In simple terms, association refers to a relationship between variables where differences in one variable are predictive of differences in another. Causation, however, involves a relationship where an alteration or manipulation of one variable induces a change in another.

Measures of association are based on the probability distribution of observed random variables. A *random variable* represents a numerical outcome of a random process, experiment, or sampling procedure, with values that can vary from one individual to another. While measures of association describe how variables relate to one another, they do not necessarily imply any causal relationship between them. For example, consider a random variable measured for an individual i selected from a well-defined target population, denoted by Y_i. The conditional probability that this random variable is equal to a specific value y, given that another random variable, D_i, is equal to a specific value d, can be expressed as follows:

$$P\left(Y_i = y | D_i = d\right) = \frac{P\left(Y_i = y, D_i = d\right)}{P\left(D_i = d\right)}. \tag{2.1}$$

In this expression, $P\left(Y_i = y | D_i = d\right)$ represents the conditional probability that $Y_i = y$ given that $D_i = d$, $P\left(Y_i = y, D_i = d\right)$ denotes the joint probability that $Y_i = y$ and $D_i = d$, and $P\left(D_i = d\right)$ is the marginal probability that $D_i = d$. To simplify our notation, we will henceforth express these probabilities more concisely as $P\left(y|d\right) = {}^{P(y,d)}/_{P(d)}$. In general, we use capital letters to denote random variables and lowercase letters to denote specific values of these variables. By comparing $P\left(y|d\right)$ across different values of d, we can observe how the probability that $Y_i = y$ varies across levels of the random variable D_i. These differences, however, may not reflect how the probability that $Y_i = y$ would change if D_i were manipulated or altered.

If the probability of Y_i does not vary across levels of D_i, then the two variables are said to be *statistically independent*. In this situation, $P\left(y|d\right) = P\left(y\right)$, and thus Equation 2.1 can be rewritten as $P\left(y\right) P\left(d\right) = P\left(y, d\right)$, after multiplying both sides by $P\left(d\right)$. This shows that when two random variables, Y_i and D_i, are statistically independent, the joint probability that $Y_i = y$ and $D_i = d$ both occur together is simply the product of their marginal probabilities. Conversely, when these variables are not independent, meaning that they are associated with one another, then $P(y)P(d) \neq P(y, d)$.

The conditional expected value of a random variable Y_i, given that the variable D_i is equal to d, is another important measure of association. This conditional expected value can be expressed as follows:

$$\mathbb{E}[Y_i|d] = \sum_y yP(y|d). \tag{2.2}$$

It represents a probability-weighted average of the random variable Y_i given that another random variable D_i is equal to a specific value d. By comparing $\mathbb{E}[Y_i|d]$ across different values of d, we can observe how the average value of Y_i varies across levels of D_i. When the two variables are statistically independent, such that $P(y|d) = P(y)$, the conditional expected value $\mathbb{E}[Y_i|d]$ is equal to the marginal expected value $\mathbb{E}[Y_i]$, where $\mathbb{E}[Y_i] = \sum_y yP(y)$. This indicates that when Y_i and D_i are independent, the average value of Y_i does not differ across levels of D_i and is instead equal to its overall average among the entire target population.

To illustrate, consider our motivating example based on the NLSY, where Y_i represents an individual's score on the Center for Epidemiological Studies-Depression Scale (CES-D; Radloff 1977) at age 40, and D_i indicates whether or not the individual attended college before the age of 22. In this context, the conditional probability $P(y|d)$ describes the likelihood of observing different CES-D scores, separately among individuals who attended college and among those who did not. Similarly, the conditional expected value $\mathbb{E}[Y_i|d]$ captures the average CES-D score within each of these two subpopulations. If the expected value and/or the conditional probability of scores on the CES-D differs between individuals who attended college and those who did not—perhaps because these scores are generally higher, indicating a greater incidence of depressive symptoms, among those with lower levels of education—we would conclude that the two variables are associated. In other words, attending college appears to be related to, or predictive of, differences in depressive symptoms, as measured by CES-D scores.

2.2 Measures of Causation

While the conditional probabilities and expectations defined in the previous section are informative about associations between variables, our current notation does not yet offer a way to clearly define measures of causation. Associational measures reflect what the researcher observes in the world without any alteration, manipulation, or intervention. These measures address questions like, "If we observe that an individual attended college, are they more (or less) likely to be depressed?" Such questions refer to relationships in the world as it exists, and do not necessarily imply a causal relationship, only a conjunction of events.

In contrast, measures of causation go beyond merely describing what the researcher is "seeing" and instead capture the consequences of "doing" something—of intervening, manipulating, or altering the world in some way (Pearl 2009, 2010). Causal measures address questions like, "If an individual had attended college rather than not, would they have been more (or less) likely to suffer depression later on?" or "If we sent an individual to college who would not otherwise attend, would their mental health improve (or deteriorate) as a result?" These questions involve a comparison of hypothetical, counterfactual states of the world. Measures of causation are distinct from measures of association in that they are expressly designed to capture relationships involving counterfactuals, rather than merely describe observed relationships among variables in the world as it is. Potential outcomes, which we will explore next, provide a conceptual apparatus for defining measures of causation and clearly distinguishing them from measures of association.

2.2.1 Potential Outcomes

Potential outcomes offer a framework for conceptualizing causal effects and distinguishing them from measures of association. If we define the variable Y_i as the "outcome" of interest and D_i as the "exposure," then a *potential outcome* of the exposure, denoted by $Y_i\left(d\right)$, represents the value that the outcome would have taken for individual i had they experienced level d of the exposure, possibly contrary to fact. Similarly, $Y_i\left(d^*\right)$ denotes the value of the outcome that would arise for the same individual if they experienced some other level of the exposure, given by d^*, instead (Hernan and Robins 2020; Holland 1986; Rubin 1974).

In this framework, each individual is conceived to have a set of potential outcomes corresponding to all possible levels of the exposure. These variables provide the conceptual basis for defining causal effects, as they allow us to compare outcomes for the same individual under different conditions, including those that may not actually occur. For example, suppose the exposure of interest is binary, with $D_i = 1$ indicating exposure to a "treatment" condition and $D_i = 0$ indicating exposure to a "control" condition. In this scenario, $Y_i\left(1\right)$ represents the outcome that would occur for individual i if they received the treatment, while $Y_i\left(0\right)$ represents the outcome for the same individual if they did not. Thus, each individual is conceived as having two potential outcomes, $Y_i\left(1\right)$ and $Y_i\left(0\right)$, corresponding to the treatment and control conditions respectively.

However, only one of these potential outcomes is ever observed for any given individual. For individuals who are actually exposed to the treatment, their observed outcome Y_i corresponds to their potential outcome under treatment, $Y_i\left(1\right)$, while their potential outcome under the control condition, $Y_i\left(0\right)$, is not observed. Conversely, for individuals who are exposed to the control condition, their observed outcome Y_i corresponds to their potential outcome under control, $Y_i\left(0\right)$, while their potential outcome under treatment, $Y_i\left(1\right)$, is not observed. In general, if an individual's observed value for the exposure is equal to d, then their potential outcome $Y_i\left(d\right)$ is observed; otherwise, it remains unobserved and represents a counterfactual.

In our example based on the NLSY, the exposure D_i is coded 1 if an individual attended college before age 22, and 0 if they did not. The potential outcome $Y_i\left(1\right)$ therefore represents the CES-D score that would have arisen for individual i at age 40 if they had attended college earlier as a young adult, while the other potential outcome, $Y\left(0\right)$, represents their CES-D score had they not attended college. The observed outcome for each individual is given by $Y_i = D_iY_i\left(1\right) + \left(1 - D_i\right)Y_i\left(0\right)$, which reflects the value of the potential outcome under the exposure that they did in fact experience.

Table 2.1 illustrates the relationship between the observed and potential outcomes using a set of hypothetical (i.e., fabricated) data from the population targeted in the NLSY. The table shows that every individual in the target population, from $i = 1, 2, ..., N$, has two distinct potential outcomes: $Y_i\left(1\right)$ and $Y_i\left(0\right)$. It also shows that the observed outcome Y_i corresponds with the potential outcome of the exposure D_i that an individual actually experienced. For example, the first individual in this population $\left(i = 1\right)$ would have a CES-D score of 3 if they attended college and a score of 2 if they had not. Since this individual did, in fact, attend college, such that $D_{i=1} = 1$, their observed outcome $Y_{i=1}$ is a score of 3 as well.

2.2.2 Individual Causal Effects

In the counterfactual framework, causal effects are defined by comparing different potential outcomes for the same individual. Specifically, the *individual total effect* of an exposure D_i on an outcome Y_i for a particular individual i can be formally defined as follows:

Table 2.1: Hypothetical Data on College Attendance D_i, CES-D Scores Y_i, and the Potential Outcomes $\{Y_i(1), Y_i(0)\}$ in the NLSY Target Population.

	Potential Outcomes		Observed Data	
Individual	$Y_i(1)$	$Y_i(0)$	D_i	Y_i
$i = 1$	3	2	1	3
$i = 2$	2	4	1	2
$i = 3$	7	12	0	12
$i = 4$	4	4	0	4
$i = 5$	13	15	1	13
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i = N$	15	16	0	16

$$ITE_i(d, d^*) = Y_i(d) - Y_i(d^*), \tag{2.3}$$

where $Y_i(d)$ represents the outcome that would occur if individual i experienced level d of the exposure, and $Y_i(d^*)$ represents their outcome if they were instead exposed to level d^*. The difference between these two potential outcomes captures the causal effect of exposure to d versus d^* for individual i.

It is important to note that $ITE_i(d, d^*)$ may vary across individuals in the target population, indicating that the effect of the exposure on the outcome is not uniform. This heterogeneity in causal effects highlights the need to consider individual-level variation when analyzing causal relationships, as the impact of any given exposure can differ from one person to another.

For example, suppose again that the exposure is binary, with $D_i = 1$ denoting exposure to a "treatment" condition and $D_i = 0$ indicating exposure to a "control" condition. In this scenario, the individual total effect is given by $ITE_i(1, 0) = Y_i(1) - Y_i(0)$, which measures how the outcome would differ for a particular individual if they were exposed to treatment rather than control. For some individuals, exposure to the treatment may result in higher outcomes, such that $Y_i(1) > Y_i(0)$, while for others, it may have no impact or result in lower outcomes, such that $Y_i(1) \leq Y_i(0)$. Only when $Y_i(1) = Y_i(0)$ for everyone in the target population, such that $ITE_i(1, 0) = 0$ for all individuals $i = 1, 2, ..., N$, could we conclude that the exposure has no effect on the outcome whatsoever.

To further illustrate, consider again the hypothetical data from the NLSY target population in Table 2.2, which now includes a column showing the individual total effects, calculated as the difference between the potential outcomes. Let's revisit the first individual in this population ($i = 1$), who would have a CES-D score of 3 if they attended college and a score of 2 if they had not. The individual treatment effect for this person is $ITE_{i=1}(1, 0) = 3 - 2 = 1$, indicating that attending college would increase their depressive symptoms. In contrast, the second individual in this population ($i = 2$) would have a CES-D score of 2 if they attended college and a score of 4 if they had not. The effect of attending college for them is $ITE_{i=2}(1, 0) = 2 - 4 = -2$, indicating a reduction in depressive symptoms.

Table 2.2: Hypothetical Data on College Attendance D_i, CES-D Scores Y_i, Potential Outcomes $\{Y_i(1), Y_i(0)\}$, and Individual Total Effects $ITE_i(1,0)$ in the NLSY Target Population.

	Potential Outcomes			Observed Data	
Individual	$Y_i(1)$	$Y_i(0)$	$ITE_i(1,0)$	D_i	Y_i
$i=1$	3	2	$3-2=1$	1	3
$i=2$	2	4	$2-4=-2$	1	2
$i=3$	7	12	$7-12=-5$	0	12
$i=4$	4	4	$4-4=0$	0	4
$i=5$	13	15	$13-15=-2$	1	13
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=N$	15	16	$15-16=-1$	0	16

2.2.3 Average Causal Effects

The $ITE_i(1,0)$ provides a measure of causation at the individual level. Other measures of causation are similarly derived from individual potential outcomes but are designed to summarize these effects across the entire population. For example, the marginal probability of the potential outcome $Y_i(d)$, denoted as $P(Y_i(d)=y)$, represents the likelihood of observing outcome y if, contrary to fact, everyone in the target population had been exposed to d. This quantity differs from the conditional probability of the observed outcome given the exposure, $P(y|d)$, which only reflects the probability of outcome y among those in the population who were actually exposed to d. Comparing $P(Y_i(d)=y)$ across different values of d measures how the probability of the outcome would change if the exposure were manipulated or altered for the entire target population, reflecting causation rather than mere association.

Similarly, computing expected values of the potential outcomes, defined as $\mathbb{E}[Y_i(d)] = \sum_y yP(Y_i(d)=y)$, provides an average value of the outcome that would be observed if all individuals in the target population were exposed to d. Conceptually, this differs from the conditional expectation of the observed outcome, $\mathbb{E}[Y_i|d]$, which reflects the average outcome in the subpopulation of individuals who were actually exposed to d. These two expected values do not necessarily align, as the former captures an average outcome under a hypothetical manipulation of the exposure for the entire target population, while the later captures the average observed only among a particular subset of this population. Comparing $\mathbb{E}[Y_i(d)]$ across different values of d, rather than $\mathbb{E}[Y_i|d]$, measures how the average value of the outcome would change if the exposure were manipulated or altered for all members of the target population.

We can formalize these comparisons with the *average total effect* of the exposure D_i on the outcome Y_i, which is defined as follows:

$$\begin{aligned} ATE(d,d^*) &= \mathbb{E}[Y_i(d) - Y_i(d^*)] \\ &= \mathbb{E}[ITE_i(d,d^*)]. \end{aligned} \tag{2.4}$$

This effect represents the expected, or average, difference in the outcome that would occur if all individuals in the target population had experienced level d rather than d^* of the exposure. In other words, it is the average of all individual causal effects, contrasting exposure to d versus d^*, for each member of the population. Conceptually, it differs from a comparison of conditional means, such as $\mathbb{E}[Y_i|d] - \mathbb{E}[Y_i|d^*]$, in

Table 2.3: The Fundamental Problem of Causal Inference, as Illustrated using Hypothetical Data from the NLSY Target Population.

	Potential Outcomes			Observed Data	
Individual	$Y_i\,(1)$	$Y_i\,(0)$	$ITE_i\,(1,0)$	D_i	Y_i
$i = 1$	3	?	$3 - ? = ?$	1	3
$i = 2$	2	?	$2 - ? = ?$	1	2
$i = 3$	?	12	$? - 12 = ?$	0	12
$i = 4$	?	4	$? - 4 = ?$	0	4
$i = 5$	13	?	$13 - ? = ?$	1	13
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i = N$	?	16	$? - 16 = ?$	0	16

Note: The question marks denote data that we cannot observe in practice.

that the average total effect measures how the central tendency of the outcome would change as a result of manipulating the exposure for the entire population, rather than simply reflecting differences in means between two observed subpopulations–those actually exposed to d and those exposed to d^*. We will revisit the average total effect in greater detail in Chapter 3, but this quantity nicely illustrates how potential outcomes and individual causal effects offer both a conceptual framework and a corresponding notation for defining measures of causation and distinguishing them from measures of association. Population-level measures of causation compare distributions of the outcome under different counterfactual scenarios, whereas measures of association compare distributions between observed population subgroups.

2.2.4 The Fundamental Problem of Causal Inference

Defining causal effects using potential outcomes highlights a critical dilemma known as the "the fundamental problem of causal inference" (Holland 1986; Rubin 1974). The *fundamental problem of causal inference* stems from the fact that it is impossible to observe all the potential outcomes for any given individual at the same time. Although each individual is conceived to have a set of potential outcomes corresponding to different levels of the exposure, we can only observe one of them—specifically, the one that corresponds with the level of exposure an individual actually experienced. The remaining potential outcomes are unobserved, or counterfactual, meaning that we always lack some of the information needed to calculate causal effects directly.

Consider again a binary exposure, where the individual total effect of D_i on Y_i is defined as $ITE_i\,(1,0) = Y_i\,(1) - Y_i\,(0)$, the difference between the potential outcomes under the treatment and control conditions for individual i. For those who receive the treatment (i.e., for whom $D_i = 1$), we only observe their potential outcome under treatment, $Y_i\,(1)$, while their potential outcome under control, $Y_i\,(0)$, remains unobserved. Conversely, for those who do not receive the treatment (i.e., for whom $D_i = 0$), we observe their potential outcome under control, $Y_i\,(0)$, but not under treatment, $Y_i\,(1)$. As a result, we can never directly observe or compute the individual causal effect, $Y_i\,(1) - Y_i\,(0)$, for any member of the target population because one of these outcomes is always missing. This also implies that at the population level, we always lack some of the necessary data to compute summary measures of the potential outcomes, like expected values.

Table 2.3 illustrates the fundamental problem of causal inference using hypothetical data from the NLSY target population. The table now displays only the values for each variable that we are capable of observing in practice. For the first individual in this population ($i = 1$), who attended college ($D_{i=1} = 1$), we only observe their potential outcome under this level of the exposure ($Y_{i=1}(1) = Y_{i=1} = 3$). However, their potential outcome had they not attended college, $Y_i(0)$, remains unobserved, as does their individual total effect. Similarly, the third individual in this population ($i = 3$) did not attend college, so we observe their potential outcome under this exposure ($Y_{i=3}(0) = Y_{i=3} = 12$). But their potential outcome had they attended college, $Y_{i=3}(1)$, is not observed, and by extension, neither is their individual total effect.

To summarize, our inability to observe both potential outcomes for the same individual under different exposures constitutes the fundamental problem of causal inference. While each individual theoretically possesses a set of potential outcomes corresponding to all possible levels of the exposure, only one of these outcomes is ever observed in practice. The other potential outcomes are counterfactuals, that is, hypothetical scenarios that cannot be observed. As a result, measures of causation cannot be directly computed; they must instead be inferred from the data we can observe.

These challenges motivate the four-step approach to causal inference outlined in Chapter 1, which we apply throughout this book: (1) define, (2) identify, (3) estimate, and (4) scrutinize. The first step involves clearly defining a causal estimand—a measure of causation that represents the object of inference we aim to understand—using potential outcomes. Since measures of causation cannot be directly observed or computed, the second and third steps focus on identification and estimation. Identification involves linking the causal estimand, which is defined in terms of unobserved potential outcomes, to empirical quantities that can be observed and computed with population data. While identification connects observable and unobservable quantities in a target population, estimation uses data from a sample of the population to predict them. Lastly, because linking counterfactual and empirical quantities requires assumptions about the relationship between the observed and unobserved data, the fourth step in our approach involves critically scrutinizing these assumptions to assess whether our estimates provide valid insights into causation. In the next section, we move to the second step of this approach, where we address the fundamental problem of causal inference by linking counterfactual and empirical quantities through assumptions about how the potential outcomes relate to the observable data.

2.3 Resolving the Fundamental Problem

In the previous section, we outlined the fundamental problem of causal inference, which arises because we can never observe all potential outcomes for any given individual. Because of this problem, causal effects–defined in terms of potential outcomes–cannot be directly measured, only inferred. The challenge, then, is to determine how we can learn about causal effects from the data we do observe, which brings us to the concept of identification.

Identification here refers to the process of linking associational quantities—those that describe observed relationships in the population—with causal quantities—those that describe the effects we would observe if we could alter the world in a particular way. Achieving this linkage requires making certain assumptions about how the unobserved, counterfactual outcomes relate to data that can be observed (Hernan and Robins 2020; Lundberg et al. 2021).

In general, there are two different approaches to addressing the fundamental problem of causal inference: one focused on identifying individual causal effects and the other on average causal effects. Identifying individual causal effects is exceptionally challenging, as it requires very strong assumptions that are nearly impossible to justify in social science research. In contrast, while identifying average effects also relies on strong assumptions, these are comparatively weaker and can be satisfied through careful study design, particularly in randomized experiments.

2.3.1 Identifying Individual Causal Effects

Identifying individual causal effects requires strong, unverifiable assumptions about either the homogeneity of different individuals or the invariance of a single individual over time (Holland 1986). Consider a scenario where two individuals, i and j, are exposed to different conditions: individual i receives a treatment ($D_i = 1$), while individual j is exposed to a control condition ($D_j = 0$). For individual i, the observed outcome is $Y_i(1)$, while $Y_i(0)$, the potential outcome under the control condition, is unobserved. Similarly, for individual j, the observed outcome is $Y_j(0)$, while $Y_j(1)$, the potential outcome under treatment, remains unobserved. For both individuals, the fundamental problem of causal inference prevents us from directly measuring the causal effects of the exposure.

Suppose, however, that aside from their different exposures, these individuals are otherwise completely homogeneous–that is, they are identical in every way, including in their response to the exposure. Formally, this homogeneity assumption implies that $Y_i(d) = Y_j(d)$ for all values of d. In this situation, the individual effect for person i, defined as $ITE_i = Y_i(1) - Y_i(0)$, could be computed by substituting the observed outcome for person j, $Y_j = Y_j(0)$, for the unobserved potential outcome $Y_i(0)$ in the ITE_i. Similarly, the individual effect for person j, defined as $ITE_j = Y_j(1) - Y_j(0)$, could be computed by substituting the observed outcome for person i, $Y_i = Y_i(1)$, for the unobserved potential outcome $Y_j(1)$ in the ITE_j. These substitutions are possible because, under the homogeneity assumption, the potential outcomes for individuals i and j are identical, allowing us to compute the difference between observed outcomes, $Y_i - Y_j$, to recover the causal effect of exposure for each of them.

This type of homogeneity assumption is sometimes employed in the physical sciences. For example, two different molecules with the same chemical composition might be assumed to respond identically to a given manipulation in a highly controlled environment, such as a vacuum. By exposing the two identical molecules to different conditions and comparing their outcomes, the individual causal effects for each could be computed from the observed data, provided that the assumption of homogeneity holds.

Now consider a scenario in which the same individual i is exposed to both the treatment and control conditions at different points in time. At time t, the individual is exposed to the treatment ($D_{it} = 1$), and at some other time t', they are exposed to the control condition ($D_{it'} = 0$). In this case, the causal effect of the exposure at any one point in time, defined as $ITE_{it} = Y_{it}(1) - Y_{it}(0)$, cannot be measured directly for individual i. This is because, although $Y_{it}(1)$ is observed at time t, the potential outcome under the control condition, $Y_{it}(0)$, is not observed at the same time.

Suppose, however, that individual i remains temporally invariant–that is, they remain identical in every way over time, including in their response to the exposure. Furthermore, suppose that the effect of the exposure is temporally transient, meaning that exposure to treatment or control at an earlier point in time does not influence the outcome at a later point in time. Rather, the exposure only influences the outcome contemporaneously. Formally, these assumptions imply that $Y_{it}(d) = Y_{it'}(d)$ for all values of d. In this situation, the individual effect for person i at time t could be computed by comparing the observed outcomes at different points in time. Specifically, the individual total effect at time t could be computed as $Y_{it} - Y_{it'}$, since we observe $Y_{it} = Y_{it}(1)$ and can substitute $Y_{it'} = Y_{it'}(0)$ for $Y_{it}(0)$ under the assumptions of temporal invariance and exposure transience.

These types of assumptions are sometimes used in the physical sciences as well. For example, the same molecule might be observed at two different points in time, before and after a treatment is applied. If the molecule is assumed to remain invariant over time and the effect of the treatment is transient, comparing the outcomes observed at different points in time yields the individual effect of interest.

But are these assumptions ever reasonable in the social sciences? In almost all cases, the answer is no. Human populations are highly heterogeneous, both in their constitution and in their responses to different exposures. Moreover, the same individual can change significantly over time, and past conditions often influence future outcomes. This heterogeneity and temporal variability make it extremely difficult to identify individual effects, as the assumptions required are highly implausible. In addition, unlike in the physical sciences, where such assumptions might be justified in studies of inanimate matter under carefully controlled conditions, they cannot be validated or met by design in social science research. As a result, identifying individual causal effects is rarely pursued in human populations. Instead, social scientists typically focus on identifying average causal effects, which require comparatively weaker assumptions that can be satisfied by experimental design.

2.3.2 Identifying Average Causal Effects

Given the inherent challenges of identifying individual causal effects, social scientists typically focus on identifying average causal effects instead, as shifting focus to average effects allows researchers to draw causal inferences using more plausible assumptions (Holland 1986). Consider the average total effect of a binary exposure D_i on an outcome Y_i, defined as $ATE\left(1,0\right) = \mathbb{E}\left[Y_i\left(1\right) - Y_i\left(0\right)\right]$. Due to the fundamental problem of causal inference, we can never directly observe the expected values of both potential outcomes, $\mathbb{E}\left[Y_i\left(1\right)\right]$ and $\mathbb{E}\left[Y_i\left(0\right)\right]$, for the entire target population, and thus the average total effect cannot be measured directly.

Instead, it is only possible to observe $\mathbb{E}\left[Y_i\left(1\right)|D_i = 1\right]$, the expected value of the potential outcome under treatment among those who actually receive the treatment, and $\mathbb{E}\left[Y_i\left(0\right)|D_i = 0\right]$, the expected value of the potential outcome under control for those who are not treated. These two quantities are equivalent to $\mathbb{E}\left[Y_i|D_i = 1\right]$ and $\mathbb{E}\left[Y_i|D_i = 0\right]$, respectively, which represent the conditional expected values of the observed outcome among treated and untreated individuals. In general, the difference between these two conditional expected values need not equal the average total effect.

Suppose, however, that the potential outcomes $\left\{Y_i\left(1\right), Y_i\left(0\right)\right\}$ are statistically independent of an individual's observed exposure D_i. Formally, this condition can be expressed as $\left\{Y_i\left(1\right), Y_i\left(0\right)\right\} \perp D_i$, where $\perp$ denotes statistical independence. Recall that when any two variables–for example, X_i and Z_i–are statistically independent, the expected value of one variable does not differ across levels of the other. This means that when two variables are independent, their conditional expected values are equal to their marginal expected values–that is, $\mathbb{E}\left[X_i|Z_i\right] = \mathbb{E}\left[X_i\right]$ and $\mathbb{E}\left[Z_i|X_i\right] = \mathbb{E}\left[Z_i\right]$ when $X_i \perp Z_i$.

If the potential outcomes are independent of the observed exposure, then, we can substitute the conditional expected value of the observed outcome among treated individuals, $\mathbb{E}\left[Y_i|D_i = 1\right]$, for the marginal expected value of the potential outcome under treatment, $\mathbb{E}\left[Y_i\left(1\right)\right]$. This substitution is possible because $\mathbb{E}\left[Y_i\left(1\right)\right] = \mathbb{E}\left[Y_i\left(1\right)|D_i = 1\right] = \mathbb{E}\left[Y_i|D_i = 1\right]$ when $Y_i\left(1\right) \perp D_i$. Similarly, we can also substitute the conditional expected value of the observed outcome among untreated individuals, $\mathbb{E}\left[Y_i|D_i = 0\right]$, for the marginal expected value of the potential outcome under control, $\mathbb{E}\left[Y_i\left(0\right)\right]$. As before, this is because $\mathbb{E}\left[Y_i\left(0\right)\right] = \mathbb{E}\left[Y_i\left(0\right)|D_i = 0\right] = \mathbb{E}\left[Y_i|D_i = 0\right]$ when $Y_i\left(0\right) \perp D_i$. Thus, under the independence assumption, the average total effect can be expressed in terms of observable rather than counterfactual quantities:

$$\begin{aligned} ATE\left(1,0\right) &= \mathbb{E}\left[Y_i\left(1\right)-Y_i\left(0\right)\right] \\ &= \mathbb{E}\left[Y_i|D_i=1\right]-\mathbb{E}\left[Y_i|D_i=0\right]. \end{aligned} \tag{2.5}$$

In words, if the potential outcomes are independent of the exposure, the expected difference in the outcome that would occur if all individuals in the target population had been treated rather than not can be equated with a comparison of observed mean outcomes between those who actually received the treatment and those who were exposed to the control condition.

The independence assumption is commonly used to identify average total effects in the social sciences, particularly in randomized experiments (Hernan and Robins 2020; Holland 1986; Morgan and Winship 2014; Rubin 1974). In a typical experiment, the exposure of interest is randomly assigned, and then the outcome is measured sometime later on. When assignment to the treatment and control conditions is determined completely at random, the exposure D_i is statistically independent of the potential outcomes $\{Y_i\left(1\right), Y_i\left(0\right)\}$ by design. As a result, comparing the mean of the observed outcome among those assigned to the treatment versus control condition provides an appropriate estimate of the average total effect. Random assignment of the exposure ensures that the independence assumption holds, allowing average total effects defined in terms of counterfactuals to be equated with a difference in observed means.

In many cases, however, conducting a randomized experiment is not feasible, and without random assignment of the exposure, the independence assumption may not hold. If we attempt to compute the average total effect, $\mathbb{E}\left[Y_i\left(1\right)-Y_i\left(0\right)\right]$, by comparing the observed conditional means, $\mathbb{E}\left[Y_i|D_i=1\right]-\mathbb{E}\left[Y_i|D_i=0\right]$, outside of a randomized experiment, our causal inferences can be mistaken, as these two quantities will not be equivalent whenever the potential outcomes are not independent of the exposure.

Mistaken inferences about average causal effects can occur when individuals exposed to the treatment and control conditions differ in other ways that also affect the outcome. In this situation, the observed difference in mean outcomes reflects not just any possible effect of the exposure, but also the influence of these other factors. Random assignment addresses this problem by ensuring that all other factors are balanced, in expectation, between the treated and untreated groups. But without randomization, these factors may be imbalanced, making it difficult to distinguish the average causal effect of the exposure from other influences.

Suppose, however, that it were possible to measure all the other factors that influence the outcome and differ between treated and untreated individuals. Within subpopulations defined by these factors, the exposure could then be treated as if it were randomly assigned. This implies that the potential outcomes are conditionally independent of the exposure, given these other factors, which can be formally expressed as $\{Y_i\left(0\right), Y_i\left(1\right)\} \perp D_i|C_i$, where C_i represents the set of variables that affect the outcome and differ across the treated and untreated groups. In substantive terms, this assumption means that within each subpopulation defined by the different levels of C_i, the potential outcomes are independent of the exposure. While the exposure may not be random across the entire population, it is "as good as" random within each of these subpopulations (Hernan and Robins 2020; Morgan and Winship 2014).

If the potential outcomes are independent of the exposure within subpopulations defined by C_i, we can substitute the conditional expected value of the observed outcome among treated individuals with covariates C_i, denoted by $\mathbb{E}\left[Y_i|D_i=1, C_i\right]$, for the conditional expected value of the potential outcome under treatment among individuals with the same covariates, $\mathbb{E}\left[Y_i\left(1\right)|C_i\right]$. This substitution is possible is because $\mathbb{E}\left[Y_i\left(1\right)|C_i\right] = \mathbb{E}\left[Y_i\left(1\right)|D_i=1, C_i\right] = \mathbb{E}\left[Y_i|D_i=1, C_i\right]$ when $Y_i\left(1\right) \perp D_i|C_i$. Similarly, we can also substitute the conditional expected value of the observed outcome among untreated individuals with covariates C_i, denoted by $\mathbb{E}\left[Y_i|D_i=0,C_i\right]$, for the conditional expected value of the potential outcome under control among individuals with these covariates, $\mathbb{E}\left[Y_i\left(0\right)|C_i\right]$. This substitution follows from the fact that $\mathbb{E}\left[Y_i\left(0\right)|C_i\right]=\mathbb{E}\left[Y_i\left(0\right)|D_i=0,C_i\right]=\mathbb{E}\left[Y_i|D_i=0,C_i\right]$ when $Y_i\left(0\right)\perp D_i|C_i$. As a result, the average total effect within each subpopulation defined by the covariates C_i can be expressed as $\mathbb{E}\left[Y_i\left(1\right)-Y_i\left(0\right)|C_i\right]=\mathbb{E}\left[Y_i|D_i=1,C_i\right]-\mathbb{E}\left[Y_i|D_i=0,C_i\right]$, which represents the difference in the observed mean outcome between those exposed to the treatment versus control condition but who otherwise have the same values on the covariates C_i.

Then, to compute the average total effect for the entire target population, we need only apply *the law of iterated expectations*, which states that the marginal expected value of a variable is equal to the expected value of its conditional expectation. Formally, for any two variables, X_i and Z_i, the law of iterated expectations states that $\mathbb{E}\left[X_i\right]=\mathbb{E}\left[\mathbb{E}\left[X_i|Z_i\right]\right]$. Applying this law to the subpopulation effects outlined previously yields an expression for the average total effect among the entire target population that only depends on observable quantities:

$$\begin{aligned}ATE\left(1,0\right)&=\mathbb{E}\left[\mathbb{E}\left[Y_i\left(1\right)-Y_i\left(0\right)|C_i\right]\right]\\&=\mathbb{E}\left[\mathbb{E}\left[Y_i|D_i=1,C_i\right]-\mathbb{E}\left[Y_i|D_i=0,C_i\right]\right],\end{aligned}\tag{2.6}$$

where the iterated expectation in the final equality can also be written as $\sum_c\left(\mathbb{E}\left[Y_i|D_i=1,c\right]-\mathbb{E}\left[Y_i|D_i=0,c\right]\right)P\left(c\right)$. Thus, under the conditional independence assumption, we can compute the average total effect by appropriately averaging mean differences in the observed outcome between treated and untreated individuals across subpopulations defined by the covariates C_i.

Specifically, the expression in Equation 2.6 shows that, under the assumption of conditional independence, the average total effect is a weighted average of the mean difference in observed outcomes between treated and untreated individuals within each subpopulation defined by the covariates. The weights here correspond to the marginal probability that an individual belongs to each subpopulation, denoted by $P\left(c\right)$. In other words, the total effect can be computed by comparing the observed mean outcome between treated and untreated individuals within each subpopulation, and then averaging these differences across all subpopulations, with each difference weighted by the probability of subpopulation membership.

The conditional independence assumption is also widely used to identify average total effects in the social sciences, particularly in observational studies where the exposure is not randomly assigned (Hernan and Robins 2020; Holland 1986; Rubin 1974). In this type of study, researchers attempt to measure all the relevant factors that differ between treated and untreated individuals and that also influence the outcome. If they succeed, such that the exposure is conditionally independent of the potential outcomes given these factors, comparing the observed mean outcome between treated and untreated individuals within levels of the measured covariates, and then appropriately averaging these differences together, yields a valid estimate of the average total effect. However, there is no guarantee in observational studies that all relevant covariates have been measured. Some may remain unobserved, and unlike randomized experiments, observational studies provide no assurance that the conditional independence assumption is satisfied. As a result, there is always a lingering concern that causal inferences based on this approach might be mistaken. In Chapters 3 to 5, we will revisit this approach to identifying average causal effects in greater detail. For now, our goal is to offer some intuition for how the fundamental problem of causal inference can be addressed through assumptions about the relationship between observable and unobservable data.

To illustrate, consider our example from the NLSY, where the exposure D_i represents college attendance and the outcome Y_i is an individual's CES-D score at age 40. Because the NLSY is an observational study, and individuals were not randomly assigned to attend college, it is unlikely that the potential outcomes are independent of the exposure. If we were to naively compare the average CES-D scores between those who attended college and those who did not, our conclusions about the average total effect would almost certainly be flawed. This difference would likely reflect the influence of many other factors beyond college attendance. For example, individuals who attended college likely differ from those who did not in terms of their race, gender, family background, parental characteristics, and academic ability—all of which may also influence mental health later on at midlife. Thus, the observed mean difference in CES-D scores between these groups may simply capture the influence of these other factors.

A more prudent approach would be to try to measure and adjust for these other factors. If we could measure an exhaustive set of relevant covariates, we might reasonably assume that the potential outcomes are independent of the exposure within subpopulations defined by these covariates. We could then compare mean CES-D scores within levels of the covariates and average these differences together in order to draw conclusions about the average total effect. However, if we fail to measure all relevant covariates or naively adjust for irrelevant ones, our inferences may still be mistaken. Thus, careful selection of covariates and rigorous scrutiny of key assumptions are crucial for valid causal inference in observational research.

But how do we determine which covariates must be measured and adjusted, and which can or should be ignored? Under what circumstances will our identification assumptions, like conditional independence, hold or break down? To address these questions, we turn now to directed acyclic graphs.

2.4 Directed Acyclic Graphs

In this section, we introduce directed acyclic graphs (DAGs; Pearl 1995, 2009, 2010). DAGs visually represent causal models, and they follow a set of rules that map hypothesized causal relationships to measures of statistical association. They serve several key purposes. First, DAGs allow researchers to represent entire systems of causal relationships in a transparent, concise, and intuitive way. Second, they enable the derivation of testable implications for the probability distribution of observed variables, based on a hypothesized causal model. Third, DAGs can be used to depict interventions, allowing researchers to emulate the potential outcomes of different manipulations within a complex causal system. Finally, DAGs are instrumental to identification analyses, helping to determine whether and how causal effects can be linked with observed data. This is especially valuable in studies of causal mediation, where the conditions linking direct and indirect effects to empirical data are often complex.

2.4.1 Elements of a DAG

A DAG consists of *nodes*, representing variables, and *arrows*, which represent causal effects between these variables. An arrow from one variable to another signifies that the variable at the origin of the arrow directly causes the variable at the terminus, while the absence of an arrow connecting two variables signifies that there is no direct causal effect between them for any member of the target population.

Paths in a DAG represent sequences of adjacent arrows connecting different variables. A causal, or directed, path is one in which all arrows point in the same direction. By contrast, a non-causal path has arrows that do not all align directionally. Variables directly caused by some other variable are called its *children*, while those that directly cause another variable are known as its *parents.* Similarly, the *descendants*

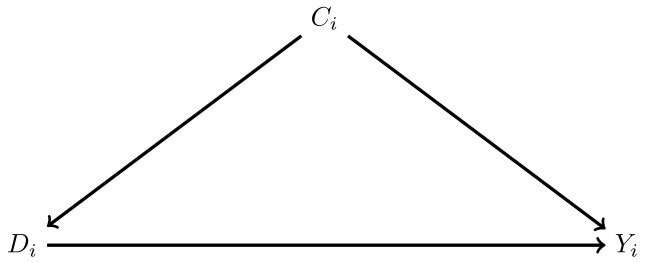


Figure 2.1: A Simple DAG with Three Variables.

of a variable in a DAG include any other variables that are directly or indirectly caused by it, such that there are directed paths connecting the focal variable to all its descendants. The *ancestors* of a focal variable, then, include any other variables that directly or indirectly cause it.

By definition, DAGs are acyclic, meaning they do not contain cycles. A *cycle* would occur if a directed path were to begin and terminate at the same variable. This acyclic property ensures that no variable can be its own descendant, ruling out the possibility of simultaneous causality and encoding the principle that causes must temporally precede their effects. Although DAGs do not permit simultaneous causality, they can still accommodate causal feedback loops by carefully specifying the temporal sequence of repeated measurements taken on the same variables over time.

To illustrate, Figure 2.1 presents a simple DAG with three variables, C_i, D_i, and Y_i, which are related in a causal system. The path $D_i \rightarrow Y_i$ is an example of a causal, or directed, path, as are the paths $C_i \rightarrow D_i$, $C_i \rightarrow Y_i$, and $C_i \rightarrow D_i \rightarrow Y_i$. The causal path involving all three variables, $C_i \rightarrow D_i \rightarrow Y_i$, is a special type of path known as a *chain*. This path links multiple variables together in a sequence: C_i affects D_i, which in turn affects Y_i. As discussed further below, chains are the defining feature of causal mediation. Based on all the causal paths in Figure 2.1, C_i is a parent of both D_i and Y_i, making D_i and Y_i its children. Similarly, D_i is a parent of Y_i, which makes Y_i is its child. The only variable with no children or other descendants is Y_i.

The DAG in Figure 2.1 also contains several non-causal paths. Specifically, the path $D_i \leftarrow C_i \rightarrow Y_i$ is a non-causal path known as a *fork*. A fork occurs when a parent (C_i) affects two children (D_i and Y_i). Conversely, the path $D_i \rightarrow Y_i \leftarrow C_i$ is another type of non-causal path referred to as an *inverted fork*. In an inverted fork, a child (Y_i) is affected by two parents (D_i and C_i). A child with two or more parents, which has multiple arrows pointing into it, is called a *collider*. Thus, inverted forks involve colliders, such as Y_i, with arrows pointing into it from both D_i and C_i (Elwert 2013; Pearl 2009).

2.4.2 DAGs as Nonparametric Causal Models

DAGs can be interpreted as nonparametric structural equation models (NPSEMs; Elwert 2013; Pearl 2009). The term "nonparametric" indicates that these models do not impose any assumptions about the probability distribution of the variables or the functional form of the causal relationships between them. The term "structural" emphasizes that the model represents the causal process generating the observed variables, rather than merely describing statistical associations. In NPSEMs, each child variable is causally determined by a generic function of its parents, along with a random disturbance that accounts for the influence of unobserved factors.

To illustrate, consider Figure 2.2, which extends the DAG discussed previously by explicitly including a

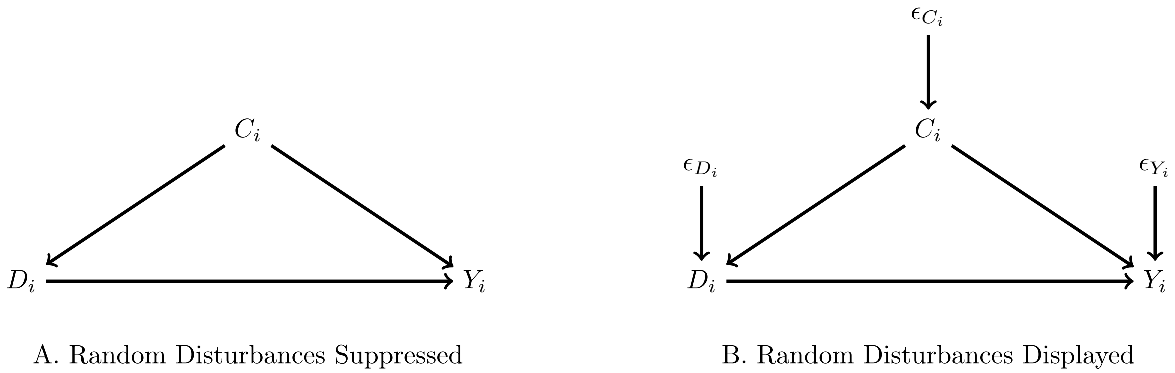


Figure 2.2: A Simple DAG with and without Random Disturbances Displayed.

Note: The $\{\epsilon_{C_i}, \epsilon_{D_i}, \epsilon_{Y_i}\}$ terms are a set of random disturbances suppressed from Panel A for visual simplicity but depicted explicitly in Panel B for illustration. The two graphs are functionally equivalent.

set of random disturbances. These disturbances represent the unobserved factors that influence each observed variable within the causal system. Although these disturbances are always implicitly present in a DAG, they are often omitted from the graph for visual simplicity, as in Figure 2.1. However, when desired, they can be included the graphical representation to provide a more complete view of the causal process underlying the observed data.

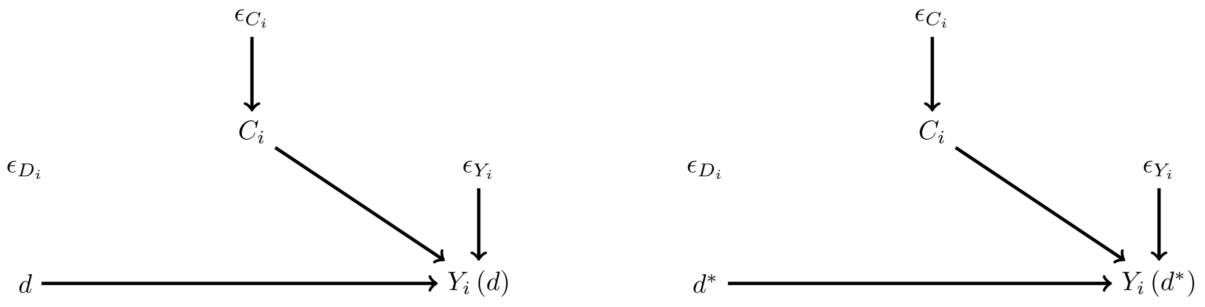
The DAG in Figure 2.2 corresponds to the following set of nonparametric structural equations:

$$\begin{aligned} C_i &:= f_{C_i}\left(\epsilon_{C_i}\right) \\ D_i &:= f_{D_i}\left(C_i, \epsilon_{D_i}\right) \\ Y_i &:= f_{Y_i}\left(C_i, D_i, \epsilon_{Y_i}\right). \end{aligned} \tag{2.7}$$

In these equations, the epsilon terms $\{\epsilon_{C_i}, \epsilon_{D_i}, \epsilon_{Y_i}\}$ represent the random disturbances, while $\{f_{C_i}, f_{D_i}, f_{Y_i}\}$ represents a set of functions that place no restrictions on the form of the causal relationships among variables or their probability distributions. The $:=$ symbol is an assignment operator that explicitly signals the direction of causal influence.

For example, the equation $Y_i := f_{Y_i}\left(C_i, D_i, \epsilon_{Y_i}\right)$ indicates that Y_i is generated as a function of C_i, D_i, and ϵ_{Y_i}, meaning that the values of Y_i are determined based on the previously realized values of its parents and a random disturbance. This equation does not specify any particular probability distribution for Y_i or its inputs, nor does it imply any specific functional form for $f_{Y_i}\left(C_i, D_i, \epsilon_{Y_i}\right)$. Thus, $Y_i := f_{Y_i}\left(C_i, D_i, \epsilon_{Y_i}\right)$ can represent a wide variety of causal models, including a conventional linear and additive SEM, like $Y_i := \beta_0 + \beta_1 C_i + \beta_2 D_i + \epsilon_{Y_i}$, but it is also compatible with far more complex specifications. The function f_{Y_i} is unrestricted and can accommodate essentially any causal relationship between Y_i and its antecedents. This flexibility also applies to f_{C_i} and f_{D_i}, as with any structural equation implied by a DAG.

In sum, DAGs can be translated into NPSEMs, where each variable in the model is an unrestricted function of its parents and a random disturbance. This approach does not impose any assumptions about the distributions of the variables or the nature of their causal relationships. It simply encodes generic patterns of causal dependence between variables.



A. $D_i := d$

B. $D_i := d^*$

Figure 2.3: Mutilated DAGs Corresponding to Different Interventions on D_i.

2.4.3 DAGs, Interventions, and Potential Outcomes

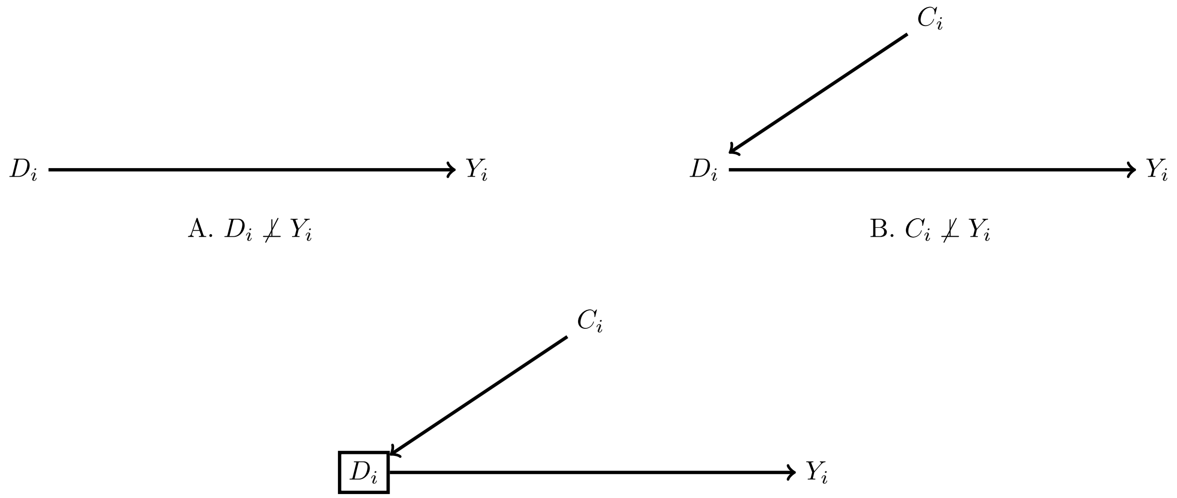
DAGs provide a graphical representation of causal relationships, while the potential outcomes framework emphasizes counterfactual reasoning, providing a notation to specify what the outcome of an exposure would be under different possible exposures. Although a typical DAG does not directly depict potential outcomes, it can still be used to represent interventions, manipulations, or alterations to the causal system and their resulting outcomes. In this way, DAGs and the potential outcomes framework are complementary.

Interventions in DAGs are represented through a process called *mutilation* (Elwert 2013; Pearl 2009). To "mutilate" a DAG, all incoming arrows to the variable(s) being intervened upon are removed. The deletion of incoming arrows reflects that, after intervention, the variable is no longer influenced by its usual causes but is instead assigned a value by a researcher or as the result of some other external influence.

For example, consider the mutilated DAG in Panel A of Figure 2.3, which illustrates an intervention where the variable D_i is set to a specific value d. In this scenario, if we intervene to assign D_i the same value for everyone, all other influences on D_i from its parents and random disturbance are completely muted, as the deterministic intervention effectively severs all the ties between D_i and its antecedent causes. Thus, as shown in Panel A of the figure, D_i has been fixed at the value d, and the incoming arrows from C_i and ϵ_{D_i} have been deleted to reflect this manipulation.

The remaining structure of the mutilated DAG allows us to trace how an intervention propagates through the system, influencing all descendant variables. Specifically, after the DAG is mutilated, the descendants of the manipulated variable now correspond to one of their potential outcomes—that is, the values that the descendants would take if their ancestor were set to a specific value. For example, after setting D_i to d in Figure 2.3, its only descendant Y_i is now equivalent to $Y_i\,(d)$, the potential outcome when D_i is set to d. Similarly, if D_i were instead set to another value d^*, its descendant Y_i would correspond to $Y_i\,(d^*)$, the potential outcome under exposure to d^*, as shown in Panel B of Figure 2.3.

Mutilating a DAG also modifies its corresponding structural equations. Before any intervention, the observed variable D_i is a function of its parents and a random disturbance. However, after the intervention, D_i is no longer influenced by these causes. Instead, it is assigned specific value, which is then treated as a fixed input in the equations governing any descendants of D_i. The resulting structural equations now reflect the potential outcomes of these descendants, capturing how the data are altered under the specified intervention. For example, under an intervention where D_i is set to d^* for everyone, as in Panel B of Figure



A. $D_i \not\perp Y_i$

B. $C_i \not\perp Y_i$

C. $C_i \perp Y_i | D_i$

Figure 2.4: Association and Independence due to Causal Paths between Variables.

Note: A box around a variable is used to denote that it has been conditioned on.

2.3, the corresponding structural equations would be expressed as follows:

$$\begin{aligned} C_i &:= f_{C_i}\left(\epsilon_{C_i}\right) \\ D_i &:= d^* \\ Y_i\left(d^*\right) &:= f_{Y_i}\left(C_i, d^*, \epsilon_{Y_i}\right), \end{aligned} \tag{2.8}$$

where $Y_i\left(d^*\right) := f_{Y_i}\left(C_i, d^*, \epsilon_{Y_i}\right)$ represents the potential outcome under exposure to d^*.

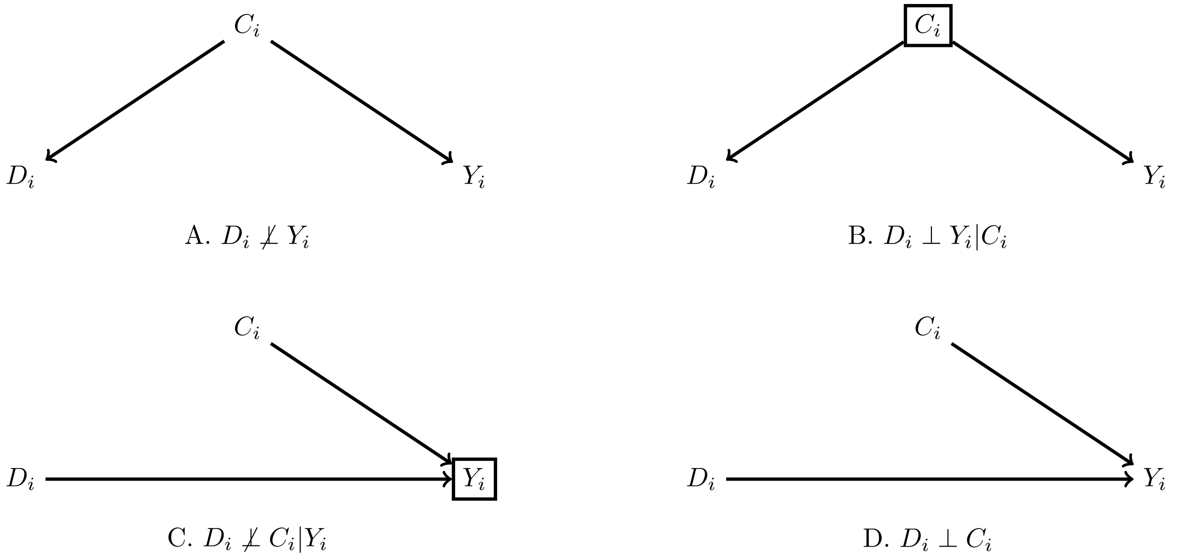
This illustrates how DAGs and the potential outcomes framework can be used together to define and reason about causal effects. DAGs provide a graphical representation of causal dependencies among variables, and when combined with the process of mutilation, they offer a convenient method for expressing how interventions alter the data generated by a system of causal relationships, revealing different potential outcomes.

2.4.4 Sources of Statistical Association in DAGs

Although DAGs help to transparently and concisely depict causal models and the outcomes of interventions, they are not merely visual aids for representing hypothesized causal relationships. They also provide a logical framework for deducing the statistical associations among observed variables that arise from a given causal model.

In DAGs, there are three fundamental sources of statistical association: causation, confounding, and endogenous selection (Elwert 2013; Elwert and Winship 2014). Each of these sources of association corresponds to a distinct structure within a graph. Causation is represented by direct causal paths and chains, while non-causal associations arise from confounding or endogenous selection, which involve forks and inverted forks, respectively.

The most straightforward source of association is causation. When one variable causally influences another, the relationship between them is represented by a directed path, or a composition of directed paths



A. $D_i \not\perp Y_i$

B. $D_i \perp Y_i | C_i$

C. $D_i \not\perp C_i | Y_i$

D. $D_i \perp C_i$

Figure 2.5: Association and Independence due to Non-causal Paths between Variables.

Note: A box around a variable is used to denote that it has been conditioned on.

forming a chain, in the DAG. For example, in Panel A of Figure 2.4, D_i directly causes Y_i, as shown by the path $D_i \to Y_i$. In this case, the observed variables D_i and Y_i would not be statistically independent in data generated from a model resembling this graph. Rather, they would be associated because D_i directly influences Y_i.

Similarly, in Panel B of Figure 2.4, C_i affects Y_i through a causal chain involving D_i, where C_i directly influences D_i, which in turn directly influences Y_i. In data generated from a model resembling this graph, C_i and Y_i would not be independent because they are linked by a chain of causal influence. However, Panel C of Figure 2.4 illustrates that C_i and Y_i would be statistically independent after conditioning on D_i, where boxes around variables in a DAG are used to indicate that they have been conditioned on. This means that C_i and Y_i would not be associated within levels of D_i, the variable linking them in a causal chain.

A second source of association is confounding, which occurs when two variables share a common cause. In a DAG, confounding is represented by a fork, where a single parent variable influences two children or descendants. For example, in Panel A of Figure 2.5, C_i directly causes both D_i and Y_i. In this situation, D_i and Y_i would not be statistically independent because they are both influenced by C_i, even though there is no direct effect or causal chain between them in the DAG. Rather, the confounding influence of C_i engenders a non-causal association between these variables. However, as shown in Panel B of Figure 2.5, conditioning on C_i eliminates the association between D_i and Y_i. In other words, within levels of C_i, D_i and Y_i would be statistically independent in data generated from a causal model resembling this graph.

The third source of association is endogenous selection, which occurs when two variables influence a common outcome, and that outcome is conditioned upon (Elwert and Winship 2014). In Panel C of Figure 2.5, C_i and D_i both affect Y_i, creating an inverted fork. On this path, Y_i is a collider, with incoming arrows originating from multiple parents. If Y_i is conditioned on, its parents C_i and D_i would become statistically associated–that is, they would not be independent within levels of the collider Y_i. However, without conditioning on Y_i, C_i and D_i would be marginally independent, as shown in Panel D of Figure 2.5. This phenomenon illustrates how conditioning on certain factors can induce associations that do not exist unconditionally, where two variables that may be marginally independent become associated within levels of their children or other descendants.

All the different sources of association in data generated from a causal model resembling a DAG can be synthesized using the concepts of d-separation and d-connection (Pearl 2009). *D-separation* is a graphical rule that determines whether a path between two variables results in a statistical association between them. A path between two variables is d-separated, or "blocked," if it contains a non-collider that has been conditioned on, or if the path contains a collider, or descendants of a collider, that have not been conditioned on. If any two variables, X_i and Z_i, are d-separated along all the paths that connect them by conditioning on another variable or set of variables, denoted by V_i, which may be empty, then X_i and Y_i are statistically independent given V_i.

D-connection is the complement of d-separation. Specifically, a path between two variables is d-connected, or "unblocked," if it is not d-separated. This implies that a path is d-connected if does not contain a non-collider that has been conditioned on or a collider that has not been conditioned on. If any two variables, X_i and Z_i, are d-connected by at least one path between them after conditioning on another variable or set of variables, denoted by V_i, which may be empty, then X_i and Y_i are not independent given V_i.

To illustrate, consider again the DAG in the lower panels of Figure 2.5, which contains a single non-causal path, the inverted fork $D_i \rightarrow Y_i \leftarrow C_i$, connecting D_i and C_i. In Panel C of the figure, the collider on this path, Y_i, has been conditioned on. As a result, D_i and C_i are d-connected, meaning that they are not independent conditional on Y_i. Conversely, in Panel D, the collider Y_i on the inverted fork connecting D_i and C_i has not been conditioned on. Thus, D_i and Y_i are d-separated along this path conditional on an empty set, meaning that they are marginally independent. The same principles and reasoning can be applied to deduce the associations and independence conditions in any other DAG.

By applying the concepts of d-separation and d-connection, DAGs can be used to determine which variables are statistically associated under different conditions, based on a particular causal model. This is especially useful for understanding how a set of causal relationships among variables translates into observable patterns of association. It also helps to determine which variables must be controlled in order to distinguish associations due to causation from those due to confounding or endogenous selection. In other words, DAGs are useful in identification analyses that assess whether and how causal effects can be linked with measures of association between observed variables.

2.4.5 Identification Analysis with DAGs

DAGs are a powerful tool for conducting identification analyses. A key challenge in causal inference is determining whether observed associations reflect a causal relationship between variables or arise from confounding or endogenous selection. DAGs provide a systematic framework for making these determinations by offering formal rules to derive statistical associations and independencies from a hypothesized causal model. In this way, they help establish the conditions under which causal effects, defined in terms of counterfactuals, can be linked to empirical quantities that can be observed and measured.

Identification is a theoretical concept that addresses whether a causal effect could, in principle, be computed exactly with an unlimited amount of data on the observed variables. It is distinct from estimation, which involves using data from a finite sample to draw inferences about a causal effect, typically accompanied by some degree of uncertainty. This distinction is important, and we will explore it further in the next chapter. Even when a causal effect is identified, practical challenges such as sampling error and model misspecification may hinder accurate estimation. However, if a causal effect is not identified, it cannot be accurately estimated, no matter how much data is collected or whether these data are modeled correctly, making identification a critical initial step in any causal analysis.

One widely used approach for conducting identification analyses with DAGs involves the *back-door criterion* (Elwert 2013; Pearl 2009). This criterion provides a simple rule for determining when the average total effect of an exposure D_i on an outcome Y_i can be identified by conditioning on another set of variables C_i, which may be empty. According to the back-door criterion, the average total effect of D_i on Y_i is identified when C_i satisfies two conditions. First, the variables in C_i must not be descendants of the exposure D_i, which ensures that conditioning on C_i does not introduce non-causal associations through endogenous selection. Second, conditioning on C_i must block all "back-door" paths—that is, paths engendering a non-causal association between D_i and Y_i due to confounding by other variables. More specifically, a back-door path is any path between D_i and Y_i that begins with an arrow pointing into D_i, indicating the presence of a potentially confounding influence. By blocking these paths, we eliminate non-causal associations due to confounding or endogenous selection, and by not adjusting for descendants of D_i, we avoid eliminating causal associations. As a result, any association between D_i and Y_i that remains after conditioning on C_i reflects a causal relationship.

To further appreciate the logic of the back-door criterion, it is helpful to consider its connection to the potential outcomes framework. When the back-door criterion for the average total effect of D_i on Y_i is satisfied by a set of variables C_i, it ensures that the exposure D_i is independent of the potential outcomes, conditional on C_i. This allows us to link the average total effect with an empirical quantity based only on the observed variables. Informally, conditional independence from the potential outcomes means that we can analyze the relationship between D_i and Y_i as if the exposure had been randomly assigned within subpopulations defined by C_i. In this situation, the average total effect can be expressed as the difference in expected outcomes among groups with different levels of the exposure in each subpopulation defined by C_i, averaged together according to the relative size of these subpopulations, as outlined in Section 2.3.2. The back-door criterion is another bridge between DAGs and the potential outcomes framework, serving to identify the variables that must be conditioned upon to satisfy key assumptions for linking counterfactual with conditional means. Conversely, it also helps to pinpoint the conditions under which these assumptions are not met.

While DAGs offer a powerful framework for guiding identification analyses, they are not without limitations in practice. The main limitation is that the true structure of the causal system underlying the observed data is rarely known with certainty. In some cases, we have partial knowledge of the true DAG, as in randomized experiments where the researcher controls the structural equation for the exposure. However, in many applications, the true causal structure is unknown, and the DAG hypothesized by the researcher may be incorrect or incomplete. If the hypothesized DAG is incorrect, any conclusions drawn from an identification analysis based thereon may also be flawed. Thus, even when researchers construct DAGs using the best available knowledge and prior evidence, they must carefully consider the possibility that the hypothesized DAG could be wrong–and that key identification assumptions might be violated–in order to ensure the robustness of their conclusions.

2.5 Direct and Indirect Causation

Causal mediation refers to the process by which an exposure affects an outcome through one or more intermediary variables, known as mediators. These mediators form part of a causal chain that links the

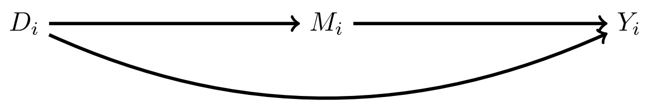


Figure 2.6: A Simple Mediation Model.

Note: D_i denotes the exposure, M_i denotes a mediator, and Y_i denotes the outcome.

exposure to the outcome. For example, in the NLSY, attending college might reduce the likelihood of unemployment, which could then reduce depression later in adulthood. In this scenario, the effect of college attendance on depression would be mediated by unemployment, acting as a link in a causal chain from education in early adulthood to mental health at midlife.

The influence of the exposure on the outcome that flows through a mediator of interest is referred to as an indirect effect. It represents part of the total effect that operates through this intermediate variable. In contrast, the influence of the exposure on the outcome that bypasses the mediator entirely is referred to as a direct effect. It represents part of the total effect that operates independently of the mediator.

In this section, we demonstrate how potential outcomes and DAGs can be used to distinguish between these forms of direct and indirect causation. Here, potential outcomes allow us to consider what the outcome would be under different counterfactual scenarios involving manipulations of both the exposure and mediator. These counterfactuals isolate the influence of the exposure that operates through the mediator from the influence that bypasses it. With DAGs, we can visually represent these distinct forms of causation, distinguishing between causal paths that include the mediator and those that do not.

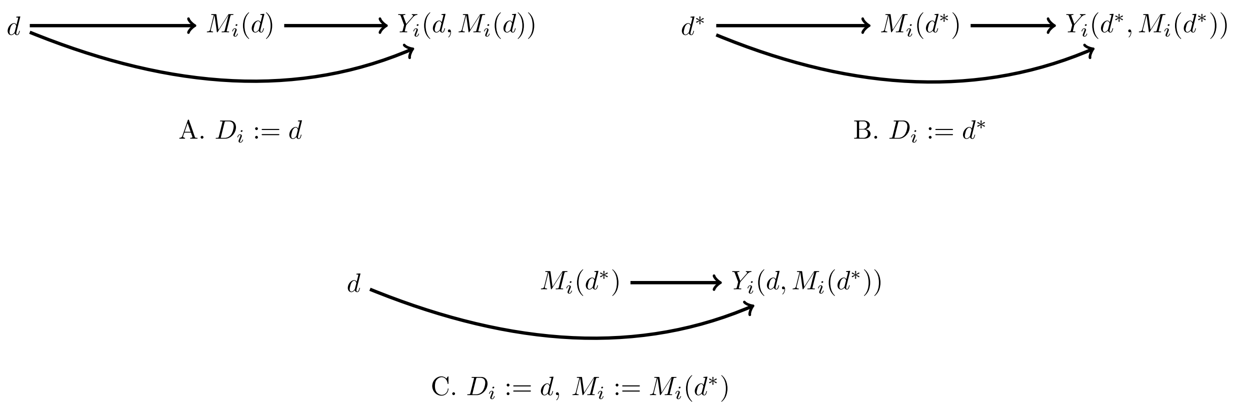
After conceptually defining direct and indirect causation using DAGs and potential outcomes, we introduce the "fundamental problem of causal mediation." This problem resembles the fundamental problem of causal inference but is even more difficult to resolve. We then discuss several related challenges that stem from this problem, setting the stage for the methods discussed in detail throughout the remainder of the book.

2.5.1 A Simple Model of Mediation

DAGs provide a clear and effective way to visually represent direct and indirect causal relationships. Figure 2.6 displays a simple mediation model using a DAG with three key variables: an exposure D_i, a mediator M_i, and an outcome Y_i.

The arrows in this graph show that the exposure D_i affects the mediator M_i, which in turn affects the outcome Y_i. This causal chain, represented by the path $D_i \to M_i \to Y_i$, captures the indirect effect of the exposure on the outcome, where the influence of D_i on Y_i is transmitted through the mediator M_i. The model also accounts for the possibility that the exposure may influence the outcome directly, independent of the mediator. This direct effect is represented by a separate arrow from D_i straight to Y_i, which bypasses the mediator entirely.

Together, the direct path ($D_i \to Y_i$) and the indirect path ($D_i \to M_i \to Y_i$) compose the total effect of the exposure on the outcome, with both mechanisms operating simultaneously. Among other goals, mediation analyses seek to decompose total effects into their direct and indirect components. By separating these paths, $D_i \to M_i \to Y_i$ and $D_i \to Y_i$, researchers can unpack the causal mechanisms through which the exposure affects the outcome.



A. $D_i := d$

B. $D_i := d^*$

C. $D_i := d,\ M_i := M_i(d^*)$

Figure 2.7: Mutilated DAGs Depicting Hypothetical Interventions in a Simple Mediation Model.

2.5.2 Nested and Cross-world Potential Outcomes

DAGs are not only useful for visualizing direct and indirect causal pathways. In addition, they also offer a framework for representing interventions and tracing how their effects propagate through downstream variables. This capability helps formalize the distinction between direct and indirect causation by allowing for comparisons of different potential outcomes under specific, hypothetical interventions.

To illustrate, consider the mutilated DAG in Panel A of Figure 2.7, which depicts an intervention where the exposure D_i is set to a specific value d. After manipulating the exposure, the descendants of this variable represent potential outcomes. Specifically, once D_i is set to d, the mediator M_i is equivalent to $M_i\left(d\right)$, its potential value under exposure to d. This value, $M_i\left(d\right)$, represents what the mediator would be for individual i if they were exposed to d, possibly contrary to fact.

Similarly, the outcome in Panel A is now given by $Y_i\left(d, M_i\left(d\right)\right)$, the potential value of Y_i under exposure to d and, by extension, under the level of the mediator that would result if the individual experienced exposure d. We refer to this quantity as a *nested potential outcome* because the potential value for M_i is nested within the potential value for Y_i. An individual's nested potential outcome $Y_i\left(d, M_i\left(d\right)\right)$ is equivalent to their conventional potential outcome $Y_i\left(d\right)$, since the potential value for Y_i under exposure d is the same as the potential value for Y_i under both exposure to d and the level of the mediator that naturally follows, $M_i\left(d\right)$.

Panel B of Figure 2.7 depicts an intervention where the exposure D_i is now set to a different value d^*. After setting D_i to d^*, the mediator M_i is equivalent to $M_i\left(d^*\right)$, its potential value under exposure to d^*. And the outcome Y_i is now equivalent to $Y_i\left(d^*, M_i\left(d^*\right)\right)$, which denotes its potential value under exposure to d^* and the level of the mediator that would naturally arise for an individual if they were exposed to d^*. This nested potential outcome, $Y_i\left(d^*, M_i\left(d^*\right)\right)$, is equivalent to the conventional potential outcome $Y_i\left(d^*\right)$ for the same reasons outlined previously. Comparing $Y_i\left(d, M_i\left(d\right)\right)$ with $Y_i\left(d^*, M_i\left(d^*\right)\right)$, then, gives the individual total effect for any person i. In contrast to the conventional potential outcomes, $Y_i\left(d\right)$ and $Y_i\left(d^*\right)$, their nested counterparts allow us to trace the effects of an intervention on an outcome down through the causal chain it initiates.

Now consider the mutilated DAG in Panel C of Figure 2.7, which depicts an intervention on both the exposure and the mediator. For this intervention, the exposure is set to d, but the mediator is set to $M_i\left(d^*\right)$, the value it would take if the individual had been exposed instead to d^*. In other words, Panel C involves setting the exposure to the level used in the intervention from Panel A while simultaneously setting the mediator to the value it would have taken under the intervention from Panel B. Under this joint intervention on both the exposure and the mediator, the outcome is represented as $Y_i\left(d, M_i\left(d^*\right)\right)$, which denotes the potential value for Y_i if an individual were exposed to d but then experienced the level of the mediator, $M_i\left(d^*\right)$, that would have occurred had they instead been exposed to d^*. We refer to this type of nested potential outcome as a *cross-world potential outcome* because it combines the exposure from one intervention with a value for the mediator from an alternative intervention where the exposure is set differently.

Comparing a cross-world potential outcome, such as $Y_i\left(d, M_i\left(d^*\right)\right)$, with the other potential outcomes from Figure 2.7 helps formalize the distinction between direct and indirect causation. For example, comparing $Y_i\left(d, M_i\left(d^*\right)\right)$ with $Y_i\left(d^*, M_i\left(d^*\right)\right)$ isolates a direct effect of the exposure on the outcome that bypasses the mediator. The difference between these two potential outcomes captures a direct effect by contrasting outcomes across different levels of the exposure–d versus d^*–while holding the mediator constant at its value under exposure to d^*. This comparison isolates a form of direct causation at the individual level.

Alternatively, comparing $Y_i\left(d, M_i\left(d\right)\right)$ with $Y_i\left(d, M_i\left(d^*\right)\right)$ isolates an indirect effect of the exposure on the outcome that operates through a causal chain involving the mediator. The difference between these potential outcomes captures an indirect effect by holding the exposure constant at d while comparing outcomes across differences in the mediator–$M_i\left(d\right)$ versus $M_i\left(d^*\right)$–that would arise naturally under exposure to d rather than d^*. By contrasting outcomes across differences in the mediator that are induced by changes in the exposure, this comparison isolates a form of indirect causation at the individual level.

Computing and comparing expected values of the nested and cross-world potential outcomes would yield a set of average effects that capture direct and indirect causation at the population level. We discuss these types of effects in greater detail in the next chapter. For now, our aim is merely to introduce the concepts of nested and cross-world potential outcomes and to provide some intuition about how they can be used to define direct versus indirect effects.

In sum, total effects are defined in terms of the potential outcomes of interventions like those depicted in Panels A and B of Figure 2.7. In contrast, direct effects are based on comparing cross-world potential outcomes, such as those in Panel C, with the potential outcomes in Panel B, while indirect effects are based on contrasts between the outcomes in Panels A and C. The comparison between outcomes in Panels C and B essentially functions to "deactivate" the causal path operating through the mediator, thereby isolating the direct influence of the exposure on the outcome for each individual. Conversely, the contrast between outcomes in Panels A and C functions to "deactivate" the direct influence of the exposure, leaving only the influence that operates through a causal chain emanating from the exposure, passing through the mediator, and terminating at the outcome (Nguyen et al. 2022).

To illustrate, Table 2.4 presents hypothetical data from the population targeted by the NLSY. In this table, the observed data consist of three variables: D_i denotes whether an individual attended college; M_i represents a potential mediator–unemployment status–coded 1 if an individual experienced a spell of unemployment between age 35 to 39, and 0 otherwise; and Y_i represents an individual's CES-D score.

The table also displays a set of potential outcomes for each individual. In the first two columns, it shows the potential values of the mediator under different exposures: $M_i\left(1\right)$ represents an individual's unemployment status if they had attended college, while $M_i\left(0\right)$ denotes their unemployment status if they had not attended college. The observed mediator M_i corresponds with its potential value under the actual exposure D_i that an individual experienced. For example, the third individual in this population ($i = 3$) would have experienced a spell of unemployment if they had not attended college, but would not have been

Table 2.4: Hypothetical Data on College Attendance D_i, Unemployment M_i, CES-D Scores Y_i, and the Potential Outcomes $\{M_i(1), M_i(0), Y_i(1, M_i(1)), Y_i(0, M_i(0)), Y_i(1, M_i(0))\}$ in the NLSY Target Population.

	Potential Outcomes					Observed Data		
Individual	$M_i(1)$	$M_i(0)$	$Y_i(1) =$ $Y_i(1, M_i(1))$	$Y_i(0) =$ $Y_i(0, M_i(0))$	$Y_i(1, M_i(0))$	D_i	M_i	Y_i
$i = 1$	0	1	3	2	2	1	0	3
$i = 2$	0	0	2	4	2	1	0	2
$i = 3$	0	1	7	12	10	0	1	12
$i = 4$	1	1	4	4	4	0	1	4
$i = 5$	0	0	13	15	13	1	0	13
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i = N$	0	0	15	16	15	0	0	16

unemployed had they attended college. Since this individual did not, in fact, attend college ($D_{i=3} = 0$), their observed mediator $M_{i=3}$ is coded 1, indicating they were unemployed at midlife.

In addition, the table displays each individual's nested potential outcomes, $Y_i(1, M_i(1))$ and $Y_i(0, M_i(0))$, which correspond to their conventional potential outcomes, $Y_i(1)$ and $Y_i(0)$, as reported previously in Tables 2.1 and 2.2. The observed outcome Y_i corresponds with one of these two potential outcomes, depending on the exposure D_i and mediator M_i that an individual actually experienced. For example, the third individual in this population ($i = 3$) has an observed CES-D score of $Y_{i=3} = 12$, which corresponds to their potential outcome $Y_{i=3}(0, M_{i=3}(0))$, because this individual did not attend college ($D_{i=3} = 0$) and, consequently, experienced a spell of unemployment ($M_{i=3}(0) = M_{i=3} = 1$). Comparing $Y_{i=3}(1, M_{i=3}(1))$ with $Y_{i=3}(0, M_{i=3}(0))$ yields the individual total effect for person $i = 3$, showing that attending college would have reduced their CES-D score by 5 points ($ITE_{i=3}(1, 0) = 7 - 12 = -5$).

Finally, Table 2.4 includes a column showing each individual's cross-world potential outcome, $Y_i(1, M_i(0))$, which represents their CES-D score if they had attended college but then experienced the level of unemployment that would have occurred for them had they not attended college. For individual $i = 3$, their cross-world potential outcome is $Y_{i=3}(1, M_{i=3}(0)) = 10$, indicating a moderate level of depressive symptoms on the CES-D scale. Comparing this outcome with $Y_{i=3}(0, M_{i=3}(0))$ yields a direct effect of college attendance: $Y_{i=3}(1, M_{i=3}(0)) - Y_{i=3}(0, M_{i=3}(0)) = 10 - 12 = -2$. Thus, for individual $i = 3$, attending college but remaining unemployed–like they did as a result of not attending college–would still reduce their CES-D score by 2 points.

Similarly, we can calculate the indirect effect for individual $i = 3$ by comparing $Y_{i=3}(1, M_{i=3}(1))$ with the cross-world potential outcome: $Y_{i=3}(1, M_{i=3}(1)) - Y_{i=3}(1, M_{i=3}(0)) = 7 - 10 = -3$. This indicates that if individual $i = 3$ had attended college and then remained employed, like they would if they attended college, instead of suffering a spell of unemployment, as they did by not attending college, their CES-D score would decline by 3 points. In other words, the change in employment status–from unemployed to employed–induced by attending college would reduce depressive symptoms for this individual.

Thus, the total effect for individual $i = 3$ is partly due to a direct influence of attending college and partly due to an indirect influence arising because attending college would have prevented a mentally taxing spell of unemployment. Similar calculations can be performed to determine the direct and indirect influences of

Table 2.5: The Fundamental Problem of Causal Mediation, as Illustrated using Hypothetical Data from the NLSY Target Population.

	Potential Outcomes					Observed Data		
Individual	$M_i(1)$	$M_i(0)$	$Y_i(1) =$ $Y_i(1, M_i(1))$	$Y_i(0) =$ $Y_i(0, M_i(0))$	$Y_i(1, M_i(0))$	D_i	M_i	Y_i
$i = 1$	0	?	3	?	?	1	0	3
$i = 2$	0	?	2	?	?	1	0	2
$i = 3$	?	1	?	12	?	0	1	12
$i = 4$	?	1	?	4	?	0	1	4
$i = 5$	0	?	13	?	?	1	0	13
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i = N$	?	0	?	16	?	0	0	16

Note: The question marks denote data that we cannot observe in practice.

college attendance on depression for every member of the target population.

2.5.3 The Fundamental Problem of Causal Mediation

Using nested and cross-world potential outcomes to define direct and indirect effects brings to light another daunting challenge for causal inference, particularly when drawing inferences about causal mediation. As discussed in Section 2.2.4, the fundamental problem of causal inference is that we cannot simultaneously observe all the potential outcomes that define individual total effects (Holland 1986; Rubin 1974). We can only observe the potential outcome corresponding to the actual exposure an individual experienced, while the other potential outcomes remain unobserved, or counterfactual. Therefore, identifying total effects requires assumptions about these unobserved data.

The difficulties associated with identifying direct and indirect causation are even greater. While the potential outcome $Y_i(d) = Y_i(d, M_i(d))$ is observed for individuals with $D_i = d$, and the alternative potential outcome used to define total effects, $Y_i(d^*) = Y_i(d^*, M_i(d^*))$, is observed for individuals with $D_i = d^*$, the cross-world potential outcome used to define direct and indirect effects, $Y_i(d, M_i(d^*))$, is never observed for anyone! We refer to this missing data problem as *the fundamental problem of causal mediation.* Like the fundamental problem of causal inference, the fundamental problem of causal mediation makes it impossible to identify direct and indirect effects without assumptions. Moreover, because this missing data problem is more severe than in the case of total effects, the assumptions needed to resolve it tend to be much stronger and harder to satisfy, even when focusing on average effects and employing a randomized experimental design (Glynn 2021; Imai et al. 2013).

Table 2.5 illustrates the fundamental problem of causal mediation using hypothetical data from the NLSY target population. The last three columns of the table display the observed values for the exposure (college attendance), mediator (unemployment status), and outcome (CES-D scores), which can actually be measured in practice. The first set of columns presents the potential values for the mediator and outcome under different conditions.

Consider individual $i = 1$, who attended college ($D_{i=1} = 1$). For this individual, we only observe the potential value for the mediator under college attendance, the exposure they actually experienced, such that

$M_{i=1}(1) = M_{i=1} = 0$. Similarly, we only observe their potential value for the outcome under this same exposure, where $Y_{i=1}(1) = Y_{i=1}(1, M_{i=1}(1)) = Y_{i=1} = 3$. However, their potential values for the mediator and outcome had they not attended college, $M_{i=1}(0)$ and $Y_{i=1}(0) = Y_{i=1}(0, M_{i=1}(0))$, remain unobserved. As in Section 2.2.4, this reflects the fundamental problem of causal inference.

The fundamental problem of causal mediation is illustrated by the fifth column of Table 2.5, which displays the cross-world potential outcomes for each individual. It shows that for individuals who attended college ($D_i = 1$), we never observe $M_i(0)$, so we cannot observe their cross-world potential outcome, $Y_i(1, M_i(0))$, either. Conversely, for individuals who did not attend college ($D_i = 0$), we observe $M_i(0)$, but we still cannot observe their cross-world potential outcome $Y_i(1, M_i(0))$. This is because we cannot know what their outcome would have been had they attended college, contrary to fact, while experiencing the level of unemployment status that did, in fact, result from them not attending college in the world as it exists. Thus, the cross-world potential outcome is never observed for any individual in the target population, and this missing information presents a significant challenge for drawing inferences about causal mediation.

As we elaborate in the remainder of the book, resolving the fundamental problem of causal mediation requires stringent assumptions that, in most cases, cannot all be satisfied by experimental design. Given the limitations of experiments for analyzing causal mediation, researchers must frequently contend with complex forms of confounding from several different sources, including baseline variables measured before the exposure and post-exposure variables measured later, each presenting distinct challenges. Furthermore, although we introduced the concepts of direct and indirect causation using a simple example with a single mediator, in reality, the effect of an exposure on an outcome often unfolds through a web of causal pathways involving multiple mediators. In such cases, the fundamental problem of causal mediation becomes even more complex, and resolving it becomes all the more difficult.

In the chapters that follow, we address these challenges one by one. First, we examine mediation analysis in the presence of confounding by baseline variables. Next, we explore the complexities introduced by confounding due to post-exposure variables, before turning to analyses of direct and indirect effects where multiple mediators are involved. Finally, we revisit the use of randomized experiments and quasi-experimental designs to aid in analyzing direct and indirect effects, discussing strategies that mitigate or sidestep some of the challenges stemming from the fundamental problem of causal mediation, even though they do not completely resolve it. Throughout, we rely heavily on the potential outcomes framework and causal graphs, with an emphasis on clearly defining the effects of interest, identifying them through precisely articulated assumptions, and rigorously scrutinizing these assumptions to assess the robustness of our inferences.

Chapter 3

Mediation Analysis with Baseline Confounding

Attending college is widely thought to improve mental health (Heckman et al. 2018; Hout 2012; Mirowsky 2003). For example, many studies have documented an inverse association between post-secondary education and depression (Yan and Williams 1999), and there is considerable evidence that this association is causal (Adams et al. 2003; Miech et al. 1999; Warren 2009). In other words, going to college seems to reduce the likelihood of becoming depressed. But how does the causal effect of education on depression come about? What are the mechanisms, or processes, that lead to lower levels of depression among adults after they have attended college?

Education may improve mental health through several different channels. As we suggested in Chapter 2, one possibility is that a more advanced education reduces depression by protecting its recipients from financially strenuous and mentally taxing spells of unemployment. Highly educated adults often experience greater demand for their labor relative to its supply on the market, and thus they tend to have more stable and more lucrative career paths than those with lower levels of education (Card 1999; Hout 2012). Alternatively, the effect of education on mental health might not have much to do with the risk of unemployment at all. Attending college could improve mental health through other mechanisms–for example, by providing greater access to health information, by promoting a healthier lifestyle, or by cultivating social support (Lee 2011; Mirowsky 2003). How might we determine if unemployment, in particular, mediates the effect of post-secondary education on depression?

In this chapter, we introduce methods for analyzing whether a single mediator explains the effect of an exposure on an outcome in settings where any potentially confounding variables, if present, are measured prior to the exposure of interest or otherwise are not exposure-induced. To begin, we present this basic mediation model using causal graphs. Next, we outline several estimands in detail that we forshadowed previously in Chapter 2, known as natural direct and indirect effects, which together capture how the influence of an exposure on an outcome is transmitted by a mediator of interest. We then describe a set of conditions that allow us to identify and estimate these effects from observed data nonparametrically–that is, without assuming the data come from some prescribed distribution model determined by a small number of parameters and a particular functional form. Finally, we explain why estimating direct and indirect effects nonparametrically can be impossible or impractical in many situations, and we show instead how they can be estimated parametrically with linear models, with nonlinear models, and with procedures that involve re-weighting the observed data in different ways.

Throughout the chapter, we illustrate key concepts and methods by analyzing the role of unemployment in mediating the effect of college attendance on depression among a sample of adults in the 1979 National Longitudinal Survey of Youth (NLSY; Bureau of Labor Statistics 2019). We conclude with a second empirical illustration that examines whether the effect of a job training program on subsequent employment is mediated by the self-confidence that program participation might instill. Stata and R codes for implementing the analyses described in this chapter are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3`. The specific scripts and data files used to generate each table or figure are hyperlinked directly in their footnotes.

3.1 Graphical Mediation Models

A basic mediation model is depicted in Figure 3.1, which displays a directed acyclic graph (DAG; Elwert 2013; Pearl 2009). Recall that the nodes in a DAG denote random variables, and an arrow from one variable to another represents a direct causal effect. A DAG can be interpreted as a nonparametric structural equation model, that is, it summarizes a set of causal relationships between variables without specifying their functional form. Although DAGs may resemble path analytic diagrams, which have long been used to analyze mediation in the social sciences (e.g., Alwin and Hauser 1975; Duncan 1966; Goldberger and Duncan 1973), they do not encode any restriction that the causal relationships among variables must be expressed through a system of linear equations. Rather, in a DAG, arrows between variables represent causal effects of arbitrary form.

The model in Figure 3.1 is composed of three variables: an exposure D, an outcome Y, and a mediator of interest M. This DAG is identical to the one introduced at the end of the previous chapter, with the exception that we have now omitted the subscript i from each variable. Throughout the remainder of the book, we will continue to omit these individuating subscripts to simplify our notation. However, it is important to remember that all the variables we consider represent measurements taken from different individuals.

The arrows connecting the variables in Figure 3.1 define a set of causal relationships between them. The arrows emanating from D to M and from M to Y, in particular, are what make this graph a mediation model. They indicate that the exposure directly affects the mediator, which in turn directly affects the outcome. In other words, the exposure affects the outcome through a causal chain in which D first causes M and then M causes Y. Causal chains, like the $D \to M \to Y$ path, are the defining feature of mediation models. The second arrow emanating from D and terminating at Y without passing through M indicates that the exposure also affects the outcome directly. The presence of a direct causal relationship, denoted by the $D \to Y$ path, implies that any effect of the exposure on the outcome is not transmitted exclusively through the mediator of interest M. Mediation analyses aim to discover whether and to what extent the effect of an exposure D on an outcome Y operates directly ($D \to Y$) versus indirectly through a mediator ($D \to M \to Y$).

Although it clearly illustrates the central features of causal mediation, the model depicted in Figure 3.1 is overly simplistic, as it does not incorporate any other variables that may affect the exposure, mediator, or outcome. Variables that jointly affect the exposure, mediator, and outcome are of special concern because they would confound the causal relationships of interest. As outlined in Chapter 2, confounding occurs when two variables share a common cause, known as a confounder. In this situation, any association between these two variables may reflect not only the causal effect of one variable on the other but also

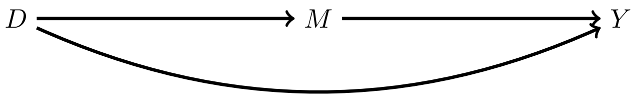


Figure 3.1: Another Simple Mediation Model.

Note: D denotes the exposure, M denotes a mediator, and Y denotes the outcome.

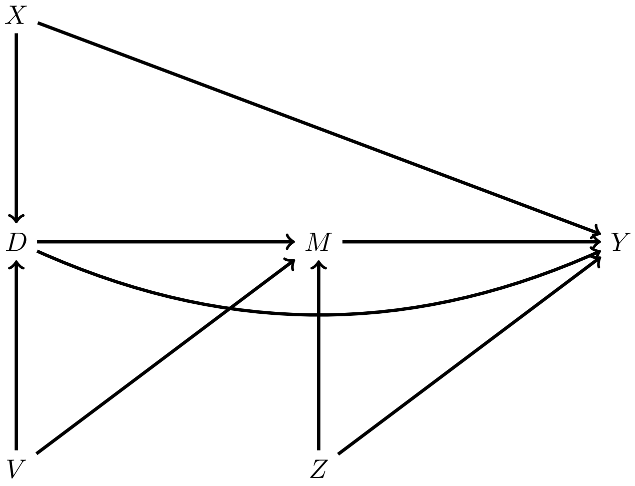


Figure 3.2: A Simple Mediation Model with Exposure-Outcome, Exposure-Mediator, and Mediator-Outcome Confounding.

Note: D denotes the exposure, M denotes a mediator, Y denotes the outcome, X denotes a set of exposure-outcome confounders, V denotes a set of exposure-mediator confounders, and Z denotes a set of mediator-outcome confounders.

the spurious covariation between them that is induced by sharing an antecedent. Because it is difficult to devise experiments that ensure the relationships between an exposure, mediator, and outcome of interest are all unconfounded by design, models of causal mediation must typically incorporate at least some type of confounding.

Figure 3.2 displays a second mediation model that includes three additional variables: X, V, and Z. These variables are not of immediate scientific interest, but we must contend with them nevertheless because they confound the causal relationships in which we are interested–namely, those connecting the exposure, mediator, and outcome. The arrows emanating from X into D and from X into Y indicate that X jointly affects both the exposure and outcome. Because it affects both D and Y, X is said to confound the effect of exposure on the outcome, and variables like X are referred to as *exposure-outcome confounders*. Similarly, the arrows emanating from V into D and from V into M indicate that V jointly affects both the exposure and the mediator. Because it affects both D and M, V is said to confound the effects of exposure on the mediator, and variables like V are referred to as *exposure-mediator confounders*. Finally, the arrows emanating from Z into M and from Z into Y indicate that the mediator and outcome are confounded by Z, and thus variables like Z are referred to as *mediator-outcome confounders*.

Arrows that are absent from a DAG are at least as important as those that are present. Another key feature of the model depicted in Figure 3.2 is that no arrow emanates from D into Z. The absence of a causal path from D into Z implies that the mediator-outcome confounder Z is not affected by the exposure

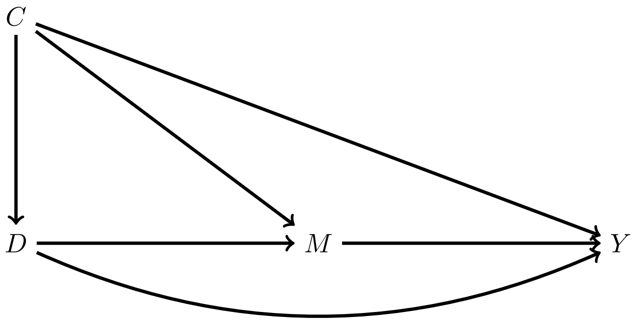


Figure 3.3: A Simple Mediation Model with Baseline Confounding.

Note: D denotes the exposure, M denotes a mediator, Y denotes the outcome, and C denotes a set of baseline confounders.

D. Thus, Z confounds the effect of the focal mediator M on the outcome Y but does not itself mediate the effect of the exposure. Because Z is not affected by D, there are not any *exposure-induced confounders* of the mediator-outcome relationship represented in this model.

Arrows from X into M, from V into Y, and from Z into D are also absent in Figure 3.2. But in the social sciences, it can be difficult to neatly partition a set of putative confounders into those that affect only the exposure and outcome (X), only the exposure and mediator (V), and only the mediator and outcome (Z). Rather, it is often the case that any confounders are thought to affect the exposure, mediator, and outcome jointly, or a model that at least allows for this possibility is preferred over a more restrictive one that precludes it.

Figure 3.3 displays a third mediation model where the different types of confounders, denoted previously by $\{X, V, Z\}$, have been replaced by a single set of confounders denoted by C. In this model, the confounders all jointly affect the exposure, mediator, and outcome of interest, as indicated by the arrows emanating from C and terminating at D, M, and Y, respectively. Thus, the variables in C confound the exposure-outcome, exposure-mediator, and mediator-outcome relationships together. As before, a key feature of this model is that no element of C is affected by the exposure D, that is, the DAG in Figure 3.3 also does not include any exposure-induced confounders. We use the term *baseline confounders* to refer to variables, like those represented by C, that may confound any of the causal relationships between D, M, and Y but are not themselves mediators transmitting an effect of exposure on the outcome.

In this chapter, we consider methods for analyzing mediation with data that arise from a causal process resembling the models depicted in either Figures 3.1, 3.2, or 3.3, where none of the confounders, if present, are exposure-induced. We tailor our presentation of these methods for models that allow general patterns of baseline confounding, as in Figure 3.3, because these are the least restrictive, but all of the approaches that we consider are also appropriate for models resembling those in Figures 3.1 or 3.2. In these settings, the baseline confounders C are either an empty set, or they subsume variables like X, V, and Z.

Whether any given data have arisen from a causal process resembling the models considered previously cannot, in general, ever be known with certainty. Rather, this must be assumed, and if the data were in fact generated by a causal process that differs from these models, then the methods we consider in this chapter may yield faulty inferences about whether and to what extent a mediator of interest transmits the causal effects of exposure on the outcome. We formalize the assumptions required for these methods to yield valid inferences about causal mediation below and then revisit them throughout the remainder of the chapter, with a special emphasis on when they might not be satisfied and on the consequences of such violations.

In our empirical illustration with the NLSY, the exposure D represents whether or not a sample member attended college before age 22, and the outcome Y represents standardized scores on the Center for Epidemiological Studies-Depression Scale (CES-D; Radloff 1977) at age 40. The mediator M represents whether or not a sample member was ever unemployed between age 35 to 39, while the baseline confounders, collectively denoted by C, include measures of race, gender, parental education, parental occupation, household size, family income, and academic achievement when a sample member was in high school. We assume that these data were generated from an underlying causal process that resembles the model depicted in Figure 3.3. In this model, college attendance may affect depression indirectly because it influences the likelihood of unemployment, as indicated by the $D \to M \to Y$ path. College attendance may also affect depression directly or through unobserved mechanisms, as indicated by the $D \to Y$ path. And finally, the whole causal process connecting education, unemployment, and depression may be confounded by the demographic characteristics of sample members and their family background, as indicated by the $D \leftarrow C \to Y$, $D \leftarrow C \to M$, and $M \leftarrow C \to Y$ paths.

3.2 Causal Estimands

In this section, we define the specific causal quantities that we use to evaluate mediation when the data are assumed to arise from a model resembling those in Section 3.1. To evaluate causal mediation in applications with a single mediator and baseline confounders only, we focus on an average total effect of the exposure on the outcome and then decompose it into the sum of a natural direct effect and a natural indirect effect (Pearl 2001; VanderWeele 2015). In this decomposition, the natural direct effect captures a component of the total effect that does not operate through the mediator of interest, while the natural indirect effect captures a component of the total effect operating through a causal chain that does involve the mediator. We also consider the controlled direct effect, which captures the influence of the exposure on the outcome if the mediator we fixed at the same value for everyone.

3.2.1 The Average Total Effect

As in Chapter 2, we define causal effects using potential outcomes notation and the counterfactual framework (Holland 1986; Rubin 1974). Let d denote a specific value of the exposure D, and let $Y(d)$ denote a *potential outcome* under exposure to d, that is, $Y(d)$ is the value of the outcome for an individual had they experienced level d of the exposure, possibly contrary to fact. In the counterfactual framework, each individual is conceived to have a set of potential outcomes corresponding to all possible values of the exposure, and contrasts between different potential outcomes define causal effects. The *average total effect* of D on Y, then, can be formally defined as follows:

$$ATE(d, d^*) = \mathbb{E}[Y(d) - Y(d^*)], \tag{3.1}$$

where $\mathbb{E}[\cdot]$ denotes an expected value and where d and d^* just denote two different values of the exposure. In words, the $ATE(d, d^*)$ is the expected, or average, difference in the outcome if individuals had experienced level d rather than d^* of the exposure.

For example, in the NLSY, the exposure D is coded 1 if a sample member attended college before age 22, and 0 otherwise. By extension, $Y(1)$ denotes a sample member's score on the CES-D at age 40 (i.e., their level of depression at midlife) had they attended college as a young adult, while $Y\left(0\right)$ denotes a sample member's level of depression had they not attended college, possibly contrary to fact. Thus, the average total effect of college attendance on depression can be expressed as $ATE\left(1,0\right)=\mathbb{E}\left[Y\left(1\right)-Y\left(0\right)\right]$, which is the expected difference in CES-D scores at age 40 if sample members had, versus had not, attended college by age 22.

3.2.2 Natural Direct and Indirect Effects

As with the potential outcomes $Y\left(d\right)$, we can also define *potential values of the mediator* under different levels of the exposure. Let $M\left(d\right)$ denote the potential value of the mediator for an individual had they experienced level d of the exposure, possibly contrary to fact. In the NLSY, $M\left(1\right)$ denotes whether a sample member would have experienced unemployment between age 35 to 39 had they previously attended college, while $M\left(0\right)$ denotes their experience with unemployment if they had not attended college.

Finally, let $Y\left(d,M\left(d\right)\right)$ denote a potential outcome under exposure to d and, by extension, under the level of the mediator that an individual would have experienced were they exposed to d. We refer to this quantity as a *nested potential outcome* because the potential value for M is nested within the potential outcome for Y. An individual's nested potential outcome $Y\left(d,M\left(d\right)\right)$ is equal to their conventional potential outcome $Y\left(d\right)$, as the potential value for Y under exposure to d is equivalent to the potential value for Y under exposure both to d and to the level of the mediator that would follow naturally from this exposure, that is, $M\left(d\right)$.

Nested potential outcomes allow us to formalize the concept of causal mediation. Specifically, using these quantities, the average total effect defined previously can be decomposed into direct and indirect components as follows:

$$\begin{aligned}
ATE\left(d,d^*\right) &= \mathbb{E}\left[Y\left(d\right)-Y\left(d^*\right)\right] \\
&= \mathbb{E}\left[Y\left(d,M\left(d\right)\right)-Y\left(d^*,M\left(d^*\right)\right)\right] \\
&= \underbrace{\mathbb{E}\left[Y\left(d,M\left(d^*\right)\right)-Y\left(d^*,M\left(d^*\right)\right)\right]}_{\text{natural direct effect}}+\underbrace{\mathbb{E}\left[Y\left(d,M\left(d\right)\right)-Y\left(d,M\left(d^*\right)\right)\right]}_{\text{natural indirect effect}}. \qquad (3.2)
\end{aligned}$$

The first equality in this decomposition just reproduces the formal definition of an average total effect in terms of the conventional potential outcomes, as in Equation 3.1 above. To arrive at the second equality, we simply substitute, or replace, the conventional potential outcomes with their corresponding nested potential outcomes. Finally, to arrive at the third equality, we both add and subtract another potential outcome, given by $Y\left(d,M\left(d^*\right)\right)$, on the right hand side of the expression and then rearrange terms. This additional potential outcome, $Y\left(d,M\left(d^*\right)\right)$, represents the value for Y if an individual were exposed to d but then were to experience the level of the mediator that would have naturally arisen for them under the alternative exposure, that is, $M\left(d^*\right)$. We refer to this type of nested potential outcome as a *cross-world potential outcome* because it involves setting the exposure at one value d and then setting the mediator at its value from an alternative counterfactual world where the individual was exposed instead to d^*.

By substituting nested for conventional potential outcomes and then by incorporating a cross-world potential outcome, we are able to decompose the total effect into the sum of two different quantities that respectively capture the concepts of direct versus indirect causation. Specifically, the first term in this decomposition is known as a *natural direct effect*:

$$NDE\left(d,d^{*}\right)=\mathbb{E}\left[Y\left(d,M\left(d^{*}\right)\right)-Y\left(d^{*},M\left(d^{*}\right)\right)\right]. \tag{3.3}$$

In words, the $NDE(d, d^*)$ is the average difference in the outcome if individuals had been exposed to d rather than d^* and if they had then experienced the level of the mediator that would have arisen naturally for them under exposure d^*, that is, $M\left(d^*\right)$. It captures an effect of the exposure on the outcome that operates through all mechanisms other than those involving the mediator of interest. The $NDE\left(d, d^*\right)$ isolates this effect by comparing outcomes across different levels of the exposure, d versus d^*, while holding the mediator constant at its value for each individual under only one level of the exposure $M\left(d^*\right)$. This comparison functions to deactivate the component of the total effect that is transmitted through a causal chain from the exposure to the mediator and from the mediator to the outcome.

The second term in this decomposition is known as a *natural indirect effect*:

$$NIE\left(d,d^{*}\right)=\mathbb{E}\left[Y\left(d,M\left(d\right)\right)-Y\left(d,M\left(d^{*}\right)\right)\right]. \tag{3.4}$$

In words, the $NIE\left(d, d^*\right)$ is the average difference in the outcome if individuals had been exposed to d and if they had then experienced the level of the mediator that would have arisen naturally for them under exposure d rather than the level of the mediator that would have arisen naturally for them under exposure d^*. It captures an effect of the exposure on the outcome that operates specifically through a causal chain involving the mediator of interest. The $NIE\left(d, d^*\right)$ isolates this effect by holding the exposure for each individual constant at d and then by comparing outcomes across differences in the mediator that would have arisen under different exposures, that is, $M\left(d\right)$ versus $M\left(d^*\right)$. This comparison functions to deactivate all causal mechanisms connecting the exposure to the outcome except for a causal chain operating through the mediator.

To summarize, the average total effect can be decomposed into a natural direct effect and a natural indirect effect, such that $ATE\left(d, d^*\right) = NDE\left(d, d^*\right) + NIE\left(d, d^*\right)$. The natural direct effect captures a component of the total effect that does not operate through a causal chain involving the mediator of interest, while the natural indirect effect captures a component of the total effect operating specifically through a casual chain from the exposure to the mediator and from the mediator to the outcome. Whether, how, and to what extent the average total effect of D on Y is mediated by M can therefore be evaluated with the $NDE\left(d, d^*\right)$ and $NIE\left(d, d^*\right)$. Are they nonzero? What is their sign and magnitude? And how does the sign and magnitude of the $NIE\left(d, d^*\right)$ compare with that of the $NDE\left(d, d^*\right)$? These are among the central questions addressed in analyses of causal mediation.

3.2.3 The Controlled Direct Effect

Natural direct and indirect effects are descriptive in that they merely characterize, or represent, the causal process by which changes in the exposure would bring about changes in the outcome via intermediate changes in a mediator. In addition to these descriptive effects, we also consider an interventional estimand known as the *controlled direct effect*. The controlled direct effect is prescriptive, rather than descriptive, in that it captures an effect of the exposure on the outcome after prescribing an intervention on the mediator that sets its value at the same level for all individuals (Acharya et al. 2016; VanderWeele 2015).

Let $Y\left(d, m\right)$ denote the potential outcome for an individual if they had experienced level d of the exposure and level m of the mediator. We refer to this type of potential outcome as a *joint potential outcome* because it is defined in terms of an intervention on both the exposure and mediator together that fixes each at some specific value for everyone. Similar to conventional potential outcomes, each individual is conceived to have a set of joint potential outcomes corresponding to all possible values of both the exposure and mediator, and in the counterfactual framework, contrasts between them define the causal effects of different interventions on both variables simultaneously.

Controlled direct effects are defined in terms of joint potential outcomes. Specifically, the controlled direct effect can be formally defined as follows:

$$CDE\left(d,d^{*},m\right)=\mathbb{E}\left[Y\left(d,m\right)-Y\left(d^{*},m\right)\right]. \tag{3.5}$$

In words, the $CDE\left(d,d^{*},m\right)$ is the average difference in the outcome if individuals were exposed to d rather d^{*} but then experienced the same level of the mediator m. When compared against the $ATE\left(d,d^{*}\right)$, the $CDE\left(d,d^{*},m\right)$ captures how the effect of the exposure on the outcome would differ after an intervention on a putative mediator that sets, or controls, its value at the same level for everyone.

The difference between natural and controlled direct effects is subtle but important. With the natural direct effect, the mediator takes on whatever value it would naturally have been for each individual under a particular level of the exposure, and this value may differ across individuals. With the controlled direct effect, by contrast, the mediator is set at the same value for every individual regardless of the exposure. This implies that the controlled direct effect may vary depending on the specific level m at which the mediator is set. It also implies that natural and controlled direct effects will not be equal except under special circumstances. In particular, these effects will only be equivalent when there is no interaction effect between the exposure and mediator on the outcome for each individual. In this scenario, the controlled direct effect would be the same regardless of the value at which the mediator is set, and by extension, $NDE\left(d,d^{*}\right)=\mathbb{E}\left[Y\left(d,M\left(d^{*}\right)\right)-Y\left(d^{*},M\left(d^{*}\right)\right)\right]$ would equal $CDE\left(d,d^{*},m\right)=\mathbb{E}\left[Y\left(d,m\right)-Y\left(d^{*},m\right)\right]$ for all levels of m. Substantively, if the effect of contrasting exposure d with d^{*} is the same no matter the value of the mediator, then it makes no difference whether the mediator for each individual is set to $M\left(d^{*}\right)$ or to some other value m, which may differ from $M\left(d^{*}\right)$.

All of the effects discussed previously have been defined independently of any model–for the outcome, mediator, or exposure, or for the joint distribution of the data as a whole. These effect definitions therefore hold regardless of the process by which the data were generated. Relatedly, they also hold regardless of whether the exposure, mediator, outcome, or any other variable in the data is continuous, ordinal, binary, or another level of measurement. They represent the most general possible definitions of total, direct, and indirect effects.

To make these definitions more concrete, however, consider our empirical example based on the NLSY. With these data, the natural direct effect can be expressed as $NDE\left(1,0\right)=\mathbb{E}\left[Y\left(1,M\left(0\right)\right)-Y\left(0,M\left(0\right)\right)\right]$, which represents the average difference in CES-D scores at age 40 if sample members had, versus had not, attended college by age 22 and if they had then experienced the level of unemployment between age 35 to 39 that would have occurred had they not attended college. The $NDE\left(1,0\right)$ captures an effect of college attendance on depression that is not due to any differences in unemployment that might have been induced by prior educational attainment.

The natural indirect effect in the NLSY can be expressed as $NIE\left(1,0\right)=\mathbb{E}\left[Y\left(1,M\left(1\right)\right)-Y\left(1,M\left(0\right)\right)\right]$, which represents the average difference in CES-D scores if sample members had attended college and then experienced the level of unemployment that would have occurred had they attended college rather than the level of unemployment that would have occurred had they not attended college. The $NIE\left(1,0\right)$ captures an effect of college attendance on depression that operates specifically through differences in unemployment induced by changes in an individual's prior educational attainment.

Finally, with a binary mediator, there are two different controlled direct effects in the NLSY, which can be expressed as $CDE\left(1,0,0\right)=\mathbb{E}\left[Y\left(1,0\right)-Y\left(0,0\right)\right]$ and $CDE\left(1,0,1\right)=\mathbb{E}\left[Y\left(1,1\right)-Y\left(0,1\right)\right]$ respectively. The first of these effects, given by $CDE\left(1,0,0\right)$, represents the average difference in CES-D scores if sample members had, versus had not, attended college and if they had not experienced any unemployment between age 35 to 39. The second, given by $CDE\left(1,0,1\right)$, is the average difference in CES-D scores if sample members had, versus had not, attended college and if they had then experienced a spell of unemployment later on between age 35 to 39. The $CDE\left(1,0,0\right)$ captures how attending college would continue to affect depression even if sample members were not to experience any unemployment at midlife, while the $CDE\left(1,0,1\right)$ captures the effect of college attendance that would persist even if sample members had all experienced a spell of unemployment.

3.3 Nonparametric Identification

The estimands defined previously involve counterfactual quantities that cannot be observed. We use the term *identification* to refer to the process by which estimands involving unobservable counterfactuals can be linked, through a series of assumptions, with observable data gathered from the entire target population. The *target population* refers to the group of individuals or set of units (e.g., schools, neighborhoods, countries) about which a researcher seeks to make inferences. In other words, it is the population to which the findings of an analysis are meant to be generalized.

More specifically, we refer to a causal estimand as identifiable under a set of assumptions if these assumptions imply that the population data are compatible with only a single value for the estimand. Conversely, a causal estimand is not identifiable when the population data are compatible with more than one value for the estimand. In other words, when a causal estimand is not identifiable, we could not discover its value even if we collected data from the entire target population, thereby obviating the error and uncertainty that arise from sampling this population as opposed to observing all its members (Lundberg et al. 2021; Hernan and Robins 2020).

An estimand can be identified in several different ways that are distinguished by the types of assumptions they invoke. *Nonparametric identification* refers to the process by which a causal estimand is linked with population data through a series of assumptions that do not involve any functional form restrictions on the probability distribution of these data. More specifically, we refer to an estimand as nonparametrically identifiable under a set of assumptions when these assumptions imply that the population data are compatible with a single value for the estimand and when they do not impose any functional form restrictions on the probability distribution of the data; otherwise, an estimand is not identifiable nonparametrically. Nonparametric identification still involves assumptions–just not assumptions about functional form–because the estimands of interest are defined in terms of counterfactuals that cannot be observed. Nevertheless, these assumptions are generally weaker than those associated with *parametric identification*, where the data are additionally assumed to come from a distribution with an explicit functional form.

In this section, we explain how the causal estimands defined previously can be nonparametrically identified. Nonparametric identification of these estimands follows a hierarchy of difficulty in that certain counterfactuals can be equated with observable data under a comparatively weaker set of assumptions than others (Nguyen et al. 2022). Identifying the expected value of conventional potential outcomes, where the exposure is set to one value and everything else follows naturally, as with the $ATE\left(d,d^{*}\right)$, requires the

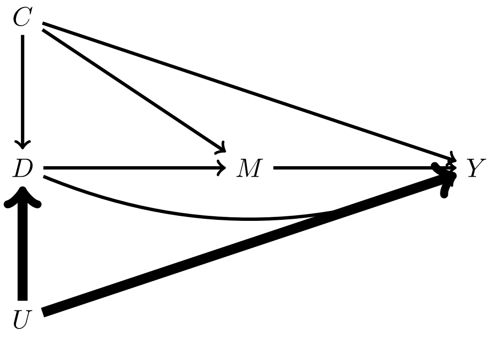


Figure 3.4: Graphical Illustration of Unobserved Exposure-outcome Confounding.

Note: D denotes the exposure, M denotes a mediator, Y denotes the outcome, C denotes a set of observed baseline confounders, and U denotes a set of unobserved confounders.

weakest assumptions, although they are still quite stringent. Identifying the expected value of joint potential outcomes, where the exposure and the mediator are both set at specific values, as with the $CDE\,(d, d^*, m)$, is more difficult and requires additional assumptions beyond those needed to identify the expected value of conventional potential outcomes. The most difficult quantities to identify are those involving cross-world potential outcomes, such as the $NDE\,(d, d^*)$ and $NIE\,(d, d^*)$, where the exposure is set to one value but the mediator is set at its value that would have arisen naturally under a different level of the exposure. Below, we explain how each of the causal effects defined in Section 3.2 can be nonparametrically identified in order of difficulty, beginning with those requiring relatively weaker assumptions and concluding with those that depend on the strongest assumptions.

3.3.1 Nonparametric Identification of Average Total Effects

The $ATE\,(d, d^*)$ can be nonparametrically identified under the following set of assumptions: (a.i) conditional independence of the exposure with respect to the potential outcomes, (a.ii) positivity of the exposure, and (a.iii) consistency of the observed with the potential outcomes (Hernan and Robins 2020; Holland 1986; Rubin 1974).

Assumption (a.i). The first of these assumptions requires that the exposure must be statistically independent of the potential outcomes conditional on the baseline confounders. It can be expressed formally as follows:

$$Y\,(d) \perp D|C, \tag{3.6}$$

where $\perp$ denotes statistical independence and all other terms are defined as before. Substantively, this assumption requires that there must not be any confounders of the exposure-outcome relationship beyond those that are measured and included in C. It would be satisfied in data generated from a process resembling the model in Figure 3.3, where there are no unobserved causes of both D and Y. It would be violated, by contrast, if the data were generated instead from a process resembling the model in Figure 3.4, where an unobserved variable U confounds the effect of D on Y. In this situation, the $ATE\,(d, d^*)$ cannot be nonparametrically identified, and thus nonparametric identification of total effects hinges on the absence of exposure-outcome confounding by unobserved variables. This condition can be met by design in experimental studies where the exposure of interest is randomly assigned because randomization ensures that the exposure is independent of all other causes of the outcome, whether they are observed or not.

Assumption (a.ii). Nonparametric identification of the $ATE\left(d, d^{*}\right)$ also requires that there must be a positive probability of all values of the exposure conditional on the baseline confounders. Formally, this assumption can be expressed as follows:

$$P\left(d \mid c\right)>0 \text { when } P\left(c\right)>0, \tag{3.7}$$

where $P\left(d \mid c\right)$ denotes the conditional probability that $D=d$ given that $C=c$ and $P\left(c\right)$ denotes the marginal probability that $C=c$. Substantively, this assumption stipulates that there must be at least some chance that individuals in the target population may experience all possible levels of the exposure within every subpopulation defined by the baseline confounders. It would be violated, for example, if individuals with a certain combination of confounder values have no chance of experiencing certain levels of the exposure. Like assumption (a.i), the positivity assumption can also be met by design in experimental studies where the exposure is randomly assigned. In this type of study, the researcher has control over the probabilities of exposure, and thus they can set them to satisfy the positivity condition. For example, in an experiment where a binary exposure is assigned based on the flip of a fair coin, the probability of receiving one versus another level of the exposure is 0.5 for everyone.

Assumption (a.iii). Finally, nonparametric identification of the $ATE\left(d, d^{*}\right)$ requires that an individual's observed outcome is consistent with their potential outcome under the exposure that they did in fact experience. This assumption can be formally expressed as follows:

$$Y=Y\left(D\right), \tag{3.8}$$

where D is an individual's observed exposure. Substantively, it requires that $Y=Y\left(d\right)$ among individuals who were in fact exposed to d. Similarly, it also requires that $Y=Y\left(d^{*}\right)$ among individuals who were in fact exposed to d^{*}.

In Chapter 2, we initially took for granted that an individual's observed outcome reflects the value of their potential outcome under the exposure that they actually experienced. However, while the consistency assumption may seem trivially true, this apparent simplicity can be deceptive. Violations of this assumption can occur when the exposure of interest is imprecisely specified or when there is interference among different individuals, meaning that the exposure experienced by one person affects the outcome of another (Hernan and Robins 2020). Relatedly, violations can arise when there are multiple versions of the exposure, each with different effects on the outcome. If there are multiple versions of an exposure, the consistency assumption requires that we clearly specify the one version used to define the potential outcomes, or alternatively, it requires that the different versions of the exposure are irrelevant because they all produce the same effects on the outcome (VanderWeele 2009a). Our consistency assumption is functionally equivalent to the "stable unit treatment value assumption" (SUTVA) described by Imbens and Rubin (2015; see also Rubin 1980), which also requires that there not be multiple versions of the exposure or any interference among individuals.

Establishing consistency between the observed and potential outcomes is relatively straightforward in randomized experiments, where the exposures of interest typically correspond to well-defined interventions or treatment protocols, and participants do not interact with each other. However, consistency can be more difficult to establish in observational studies, where exposures may not correspond with single, well-defined interventions or where interference between members of the study population is possible.

Nonparametric identification formula. If assumptions (a.i) to (a.iii) are all satisfied, the $ATE\left(d, d^{*}\right)$ can be equated with a function that is defined only in terms of observable data rather than the potential outcomes, some of which cannot be observed because they are counterfactuals. This function, which we refer to as the nonparametric identification formula for the $ATE\left(d, d^{*}\right)$, can be expressed as follows:

$$ATE\left(d, d^{*}\right)=\sum_{c}\left(\mathbb{E}[Y \mid c, d]-\mathbb{E}\left[Y \mid c, d^{*}\right]\right) P(c), \tag{3.9}$$

where $\mathbb{E}[Y|c, d]$ denotes the conditional expected value of the observed outcome Y among individuals for whom $C = c$ and $D = d$, $\mathbb{E}[Y|c, d^{*}]$ denotes the conditional expected value of the observed outcome Y among individuals for whom $C = c$ and $D = d^{*}$, and $P(c)$ denotes the marginal probability that $C = c$. In this expression, $\mathbb{E}[Y|c, d] - \mathbb{E}[Y|c, d^{*}]$ represents the difference in the expected values of the observed outcome Y between individuals who experienced different levels of the exposure, $D = d$ versus $D = d^{*}$, but who had the same values on the baseline confounders. This difference in expected values is evaluated among all the subpopulations defined by different levels of the confounders, and then these differences are averaged together, weighting each by the probability that an individual falls in a particular subpopulation. In other words, the nonparametric identification formula computes the average total effect by perfectly stratifying the target population by all levels of the confounders, evaluating the difference in the mean of the outcome between those with different levels of the exposure in each stratum, and then by computing a weighted average of all the stratum-specific differences in means, with weights equal to the relative size of each stratum. In Appendix A, we provide a step-by-step derivation of this identification formula, beginning with the formal definition of the $ATE(d, d^{*})$ in terms of counterfactuals and then invoking assumptions (a.i) to (a.iii) to arrive at the expression in Equation 3.9, which involves only the observable data $\{C, D, Y\}$ and not the potential outcomes $\{Y(d), Y(d^{*})\}$.

Consider now the nonparametric identifiability of the average total effect in our empirical example based on the NLSY. Recall that the average total effect in this example is the expected difference in CES-D scores at age 40 if individuals had, versus had not, attended college by age 22, which can be formally expressed as $ATE(1, 0) = \mathbb{E}[Y(1) - Y(0)]$ when the exposure is a binary variable denoting college attendance. This effect can be nonparametrically identified if (a.i) college attendance is conditionally independent of the potential outcomes given the baseline confounders, (a.ii) there is a positive probability that individuals attended college within each of the subpopulations defined by the baseline confounders, and (a.iii) an individual's observed score on the CES-D is consistent with their potential outcome under the exposure they did in fact experience.

Are these assumptions reasonable? In contrast to experimental studies, where the independence, positivity, and consistency assumptions can be met by design, the same assumptions are not guaranteed to hold in the NLSY, or for that matter, in any observational study, which is why causal inferences are often viewed with greater skepticism in these applications.

Assumption (a.i) here requires that there are not any variables that affect both college attendance and depression above and beyond those included in our vector of baseline confounders. Although the assumption of no exposure-outcome confounding by unobserved factors can never be verified or disconfirmed in an observational study, it is likely violated in the NLSY because there are many unmeasured factors that may affect both the chances of attending college and mental health at midlife, such as individual personality traits, childhood trauma or neglect, the experience of depression during adolescence, and so on. Thus, the $ATE(1, 0)$ is probably not identified nonparametrically, and even if the NLSY had surveyed the entire target population, the observed data would be compatible with multiple different values for this effect. In Section 3.7 below, we discuss methods for computing a range of plausible estimates for a causal effect when concerns arise about the possibility of unobserved confounding.

Assumption (a.ii) requires that there are at least some individuals who attended college, as well as some who did not, in every subpopulation defined by the baseline confounders. The positivity assumption is more reasonable in the NLSY, although it is still not completely beyond reproach. It would be violated, for example, if there are highly disadvantaged groups, such as individuals with parents who did not finish high school and who earned very low incomes, in which nobody attended college due to a lack of access or resources.

An important difference between the conditional independence and positivity assumptions is that the positivity assumption can be empirically verified. In particular, we could confirm that positivity holds in our analysis of the NLSY by stratifying the sample data across the baseline confounders and then observing that there are at least some individuals who attended college and some who did not in every stratum. Suppose, however, that there were some strata in which none of the sample members selected for the NLSY had attended college. There are two reasons why this might occur. The first is that it may have simply been due to chance. Even if the probability of attending college lies between 0 and 1 for everybody regardless of their confounders, it remains possible that we might have selected a sample where nobody attended college in certain strata because of sampling error. In this situation, positivity would hold in the target population but not in our sample from that population due to the random error inherent in any sampling procedure.

The second reason why we might not find anyone who has attended college in some of our sample strata is that there are none in the population either. In other words, individuals with certain values on the baseline confounders might never experience certain levels of the exposure, in which case positivity would not hold in the target population nor in any sample from that population. This type of structural departure from positivity seems unlikely in our analysis based on the NLSY, as an individual's educational attainment is not completely determined by their social and demographic background.

Structural versus random departures from positivity have different implications for identifiability. In the presence of structural departures from positivity, the average total effect cannot be identified nonparametrically. Even if we had data from the entire target population, there would still be certain strata of the confounders where comparing individuals with different levels of education is impossible–for example, because nobody in them attended college. In this situation, it may still be possible to nonparametrically identify a *conditional* average effect in the subpopulations where individuals actually experience all levels of the exposure, but nonparametric identification of causal effects defined over the full population is precluded by structural violations of the positivity condition (Hernan and Robins 2020).

In the presence of random departures from positivity, the average total effect can still be nonparametrically identified, but estimating these effects cannot be accomplished without invoking at least some additional assumptions about the functional form of the probability distribution from which the observed data were sampled. In this situation, a parametric model is used to "smooth over," or to "fill in," the missing information in strata with random departures from positivity by borrowing information from other strata where positivity holds in the sample data. We revisit the distinction between nonparametric and model-based estimation in Sections 3.4 and 3.5 below.

Assumption (a.iii) in the NLSY requires that, among those who did in fact complete college, an individual's observed CES-D score Y must equal their potential outcome had they completed college $Y(1)$, and conversely, among those who did not complete college, an individual's observed outcome Y must equal their potential outcome had they not completed college $Y(0)$. This assumption is also questionable, as "college attendance" is not a very precisely defined exposure and many different versions of this exposure exist in the target population. For example, some individuals attend college at highly selective universities and pursue competitive degree programs, while others attend regional public schools and pursue a less competitive course of study. These different versions of the exposure may have very different consequences for mental health at midlife. Exposures like "college attendance" that are imprecisely defined complicate the interpretation of causal effects and make it difficult to translate them into consequences of real-world interventions. How exactly might we compel someone to attend college, what type of degree would they pursue, and where?

Moreover, violations of the consistency assumption could occur if one individual's education impacts the mental health of others. For example, if an individual's social interactions with peers who did not attend college influence their own mental health outcomes, this would violate the restriction on interference that the consistency assumption implies. Similarly, shared college environments, such as group living situations or collaborative academic settings, could also lead to spillover effects, where the mental health benefits or challenges experienced by one student affect others.

In observational studies, the consistency assumption is often difficult to satisfy completely, and thus there may always be some degree of ambiguity about the interpretation of causal effects in these settings. Nevertheless, we believe that there is still value in examining causal estimands that are somewhat ambiguously defined, such as the average total effect of college attendance on depression. Although we may not be able to determine exactly what type of intervention would produce such an effect, there is value in learning, for example, that a certain amount of adult depression could have been prevented if individuals, who interact in complex ways, had been compelled, somehow, to attend a college of some sort. This knowledge might help to illuminate the etiology of depression and the role of education in preventing or producing mental health problems. It might also point toward interventions aimed at boosting educational attainment as potentially useful for improving population health, even if the exact nature of these interventions is initially unspecified and the task of identifying precise, feasible, and effective interventions is left for future research.

3.3.2 Nonparametric Identification of Controlled Direct Effects

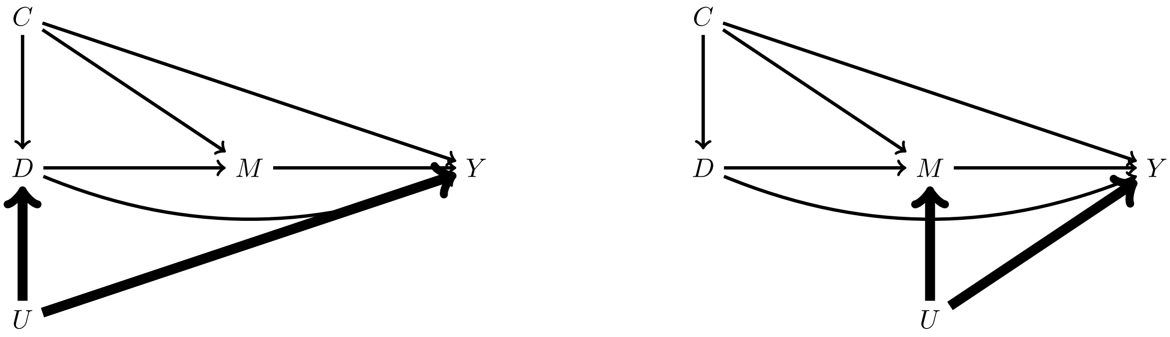
Nonparametric identification of controlled direct effects is similar to that for average total effects but requires additional and more stringent assumptions. Specifically, the $CDE\left(d, d^{*}, m\right)$ can be nonparametrically identified under the following set of assumptions: (b.i) conditional independence of the exposure with respect to the joint potential outcomes, (b.ii) conditional independence of the mediator with respect to the joint potential outcomes, (b.iii) joint positivity of both the exposure and the mediator, and (b.iv) consistency of the observed with the joint potential outcomes.

Assumption (b.i). The first of these assumptions requires that the exposure must be statistically independent of the joint potential outcomes conditional on the baseline confounders. It can be expressed formally as follows:

$$Y\left(d, m\right) \perp D | C. \tag{3.10}$$

Substantively, this assumption requires that there must not be any unobserved confounders of the exposure-outcome relationship, and thus it is essentially equivalent to the conditional independence assumption required to nonparametrically identify the average total effect, as in Section 3.3.1.

Assumption (b.ii). Nonparametric identification of the $CDE\left(d, d^{*}, m\right)$ also requires that, among individuals for whom $D = d$, the mediator must be statistically independent of the joint potential outcomes conditional on the baseline confounders. This assumption can be formally expressed as follows:



A. Unobserved Exposure-outcome Confounding

B. Unobserved Mediator-outcome Confounding

Figure 3.5: Graphical Illustration of Unobserved Exposure-outcome and Mediator-outcome Confounding.

Note: D denotes the exposure, M denotes a mediator, Y denotes the outcome, C denotes a set of observed baseline confounders, and U denotes a set of unobserved confounders.

$$Y(d,m) \perp M|C, D = d. \tag{3.11}$$

Substantively, this assumption requires that there must not be any unobserved confounders of the mediator-outcome relationship. Or in other words, aside from the confounders measured and included in C, there must not be any other factors that jointly affect both the mediator M and the outcome Y among individuals who were exposed to d.

Assumptions (b.i) and (b.ii) would both be satisfied in data generated from a process resembling the model in Figure 3.3, where there are no unobserved causes of Y that also affect either D or M. By contrast, assumption (b.i) would be violated if the data were generated instead from a process resembling the model in Panel A of Figure 3.5, where an unobserved variable U confounds the effect of D on Y. Similarly, assumption (b.ii) would be violated if the data were generated from a process resembling the model in Panel B of Figure 3.5, where the unobserved variable U now confounds the effect of M on Y. In either of these situations, the $CDE(d, d^*, m)$ cannot be nonparametrically identified, and thus nonparametric identification of controlled direct effects hinges on the absence of both exposure-outcome and mediator-outcome confounding by unobserved variables. These two conditions can be met by design in experimental studies where the exposure and mediator are jointly or sequentially randomized, that is, where individuals are first randomly assigned to different levels of the exposure and then, within levels of the exposure, they are subsequently assigned to different levels of mediator. Joint or sequential randomization ensures that both the exposure and mediator are independent of all other causes of the outcome, whether they are observed or not.

Assumption (b.iii). In addition, nonparametric identification of the $CDE(d, d^*, m)$ requires that there must be a positive probability of all values for the exposure and mediator conditional on the baseline confounders. Formally, this assumption can be expressed as follows:

$$P(d, m|c) > 0 \text{ when } P(c) > 0. \tag{3.12}$$

Substantively, it stipulates that there must be at least some chance that individuals experience all possible levels of both the exposure and mediator within every subpopulation defined by the baseline confounders.

This assumption would be violated, for example, if individuals with a certain combination of confounder values have no chance of experiencing certain levels of the exposure, certain levels of the mediator, or certain levels of the exposure and mediator together. This assumption can be met by design in experimental studies where both the exposure and mediator are randomly assigned because the researcher controls the assignment probabilities and can therefore calibrate them to satisfy this variant of the positivity condition. For example, in a sequentially randomized experiment, where assignment to different levels of a binary exposure and a binary mediator is determined by the flip of a fair coin, the probability of assignment to each of the four experimental conditions is $0.5 \times 0.5 = 0.25$ for everyone.

Assumption (b.iv). Finally, nonparametric identification of the $CDE\left(d, d^{*}, m\right)$ requires that an individual's observed outcome is consistent with their joint potential outcome under the levels of both the exposure and the mediator that they did in fact experience. This assumption can be formally expressed as follows:

$$Y = Y\left(D, M\right), \tag{3.13}$$

where D and M denote an individual's observed values on the exposure and mediator, respectively. Substantively, this assumption requires that $Y = Y\left(d, m\right)$ among individuals who were exposed to d and who experienced level m of the mediator, that $Y = Y\left(d^{*}, m\right)$ among individuals who were exposed to d^{*} and who experienced level m of the mediator, and so on for all possible values of the exposure and mediator. Similar to assumption (a.iii) in Section 3.3.1, this variant of the consistency assumption can be deceptively difficult to satisfy, especially in observational studies where violations may arise when the either the exposure or mediator are imprecisely defined, when there are multiple versions of these variables with different effects on the outcome, or when there is interference among different individuals.

Nonparametric identification formula. Nevertheless, if assumptions (b.i) to (b.iv) are all satisfied, the controlled direct effect can be expressed as a function of observable data rather than unobservable counterfactuals. This function, which we refer to as the nonparametric identification formula for the $CDE\left(d, d^{*}, m\right)$, is given by the following equation:

$$CDE\left(d, d^{*}, m\right) = \sum_{c} \left(\mathbb{E}\left[Y|c, d, m\right] - \mathbb{E}\left[Y|c, d^{*}, m\right]\right) P\left(c\right), \tag{3.14}$$

where $\mathbb{E}\left[Y|c, d, m\right]$ denotes the conditional expected value of the observed outcome Y among individuals for whom $C = c$, $D = d$, and $M = m$, where $\mathbb{E}\left[Y|c, d^{*}, m\right]$ is defined analogously, and where $P\left(c\right)$ denotes the marginal probability that $C = c$, as before. In Appendix B, we provide a step-by-step derivation of this identification formula, beginning with the formal definition of the $CDE\left(d, d^{*}, m\right)$ in terms of the joint potential outcomes and then invoking assumptions (b.i) to (b.iv) to arrive at the expression in Equation 3.14, which involves only observable data.

In this expression, $\mathbb{E}\left[Y|c, d, m\right] - \mathbb{E}\left[Y|c, d^{*}, m\right]$ denotes the difference in the expected values of the observed outcome Y between individuals who experienced different levels of the exposure, $D = d$ versus $D = d^{*}$, but the same values on both the baseline confounders and the mediator. This difference in expected values–comparing individuals who experienced different levels of the exposure but the same value of the mediator–is evaluated among all the different subpopulations defined by the baseline confounders, and then these differences are averaged together, weighting each by the probability that an individual falls in a particular subpopulation. In other words, the nonparametric identification formula in Equation 3.14 computes the controlled direct effect by perfectly stratifying the target population by all levels of the baseline confounders, evaluating the difference in the mean of the outcome between those with different levels of the exposure but the same level of the mediator within each of these strata, and then by taking a weighted average of all these stratum-specific differences in means, with weights equal to the relative size of each stratum.

Consider now the nonparametric identifiability of the $CDE\,(d, d^*, m)$ in the NLSY. Recall that there are two different controlled direct effects in this example, depending on the level to which the binary mediator is set. The first is denoted by $CDE\,(1, 0, 0)$, and it represents the expected difference in CES-D scores at age 40 if individuals had, versus had not, attended college before age 22 and if they had not experienced any unemployment between age 35 to 39. Similarly, the second effect is denoted by $CDE\,(1, 0, 1)$, and it captures the expected difference in CES-D scores if individuals had, versus had not, attended college and if they had experienced a spell of unemployment. Both effects are nonparametrically identified if (b.i) college attendance is conditionally independent of the joint potential outcomes given the baseline confounders, (b.ii) unemployment is conditionally independent of the joint potential outcomes given the baseline confounders and college attendance, (b.iii) there is a positive probability of college attendance and unemployment within each of the subpopulations defined by the baseline confounders, and (b.iv) an individual's observed score on the CES-D is consistent with their joint potential outcome under levels of the exposure and mediator that they did in fact experience. Are these assumptions reasonable in the NLSY? Probably not. Nothing guarantees that they hold in any observational study, and skepticism about whether they are satisfied in our empirical example is almost certainly warranted.

In substantive terms, assumption (b.i) requires that there must not be any factors that affect college attendance and CES-D scores above and beyond those measured and included in C, while assumption (b.ii) requires that there not be any unobserved factors that jointly affect the risk of unemployment and CES-D scores. Both assumptions are likely violated because there are many factors that are unmeasured in the NLSY and that may affect depression, educational attainment, and the risk of unemployment.

Assumption (b.iii) requires that, within every subpopulation defined by the baseline confounders, there must be at least some individuals who attended college who were employed continuously, who attended college and experienced a spell of unemployment, who did not attend college and were employed continuously, and who did not attend college and experienced unemployment. This assumption is more reasonable, as college attendance and unemployment are relatively common experiences that likely occur for at least some individuals across all socioeconomic and demographic groups in the target population.

Lastly, assumption (b.iv) requires that an individual's observed CES-D score is equal to their joint potential outcome under the combination of educational attainment and unemployment that they did in fact experience. This variant of the consistency assumption is also somewhat dubious in the NLSY because both "college attendance" and "unemployment" involve a highly diverse set of experiences with potentially variable effects on depression, and because both variables likely have spillover effects from one individual to another. Violations of the consistency assumption in this example cast a degree of ambiguity over the interpretation of controlled direct effects. In particular, they are difficult to interpret as the consequences of specific interventions on educational attainment and employment status because the exact nature of these interventions have not been precisely specified in the definition of the joint potential outcomes. Despite these challenges, there is still value in examining controlled direct effects, even in observational studies like the NLSY, provided that researchers carefully assess potential violations of their identification assumptions and qualify their inferences accordingly. We return to these issues in Section 3.7 below.

3.3.3 Nonparametric Identification of Natural Direct and Indirect Effects

Because they involve cross-world potential outcomes, nonparametric identification of natural direct and indirect effects requires the strongest assumptions. Specifically, the $NDE\left(d,d^*\right)$ and $NIE\left(d,d^*\right)$ can be nonparametrically identified under the following set of assumptions: (c.i) conditional independence of the exposure with respect to the joint potential outcomes, (c.ii) conditional independence of the mediator with respect to the joint potential outcomes, (c.iii) conditional independence of the exposure with respect to the potential values of the mediator, (c.iv) cross-world independence of the joint potential outcomes with respect to the potential values of the mediator, (c.v) joint positivity of both the exposure and mediator, and lastly, (c.vi) consistency of the observed, potential, nested potential, and joint potential outcomes.

Assumption (c.i). The first of these assumptions requires that the exposure must be statistically independent of the joint potential outcomes conditional on the baseline confounders. Substantively, this assumption requires that there must not be any unobserved factors that confound the exposure-outcome relationship. It is equivalent to assumption (b.i) in Section 3.3.1 on the identifiability of controlled direct effects, and it can be formally expressed exactly as in Equation 3.10 above.

Assumption (c.ii). Nonparametric identification of natural direct and indirect effects also requires that, among individuals for whom $D = d$, the mediator must be statistically independent of the joint potential outcomes conditional on the baseline confounders. In substantive terms, this assumption requires that there must not be any unobserved factors that confound the mediator-outcome relationship. It is equivalent to assumption (b.ii) in Section 3.3.1 on the identifiability of controlled direct effects, and it can be formally expressed exactly as in Equation 3.11 above. The conditional independence assumptions required for identifying natural direct and indirect effects therefore subsume those required for identifying controlled direct effects. In both cases, nonparametric identification hinges on the absence of unobserved variables that confound either the exposure-outcome or the mediator-outcome relationships.

Assumption (c.iii). In addition, nonparametric identification of the $NDE\left(d,d^*\right)$ and $NIE\left(d,d^*\right)$ requires that the exposure must also be statistically independent of the potential values of the mediator, conditional on the baseline confounders. Formally, this assumption can be expressed as follows:

$$M\left(d\right) \perp D|C. \tag{3.15}$$

In substantive terms, it stipulates that there must not be any unobserved confounders of the exposure-mediator relationship–that is, aside from the observed variables in C, there must not be any other factors that jointly affect both the exposure and mediator.

Assumption (c.iv). The fourth assumption required to nonparametrically identify natural direct and indirect effects is known as a “cross-world” independence assumption. This moniker comes from the fact that assumption (c.iv) involves a restriction on the relationship of the joint potential outcomes under one value of the exposure with the potential values of the mediator under a different value of the exposure. Specifically, this assumption can be expressed formally as follows:

$$Y\left(d,m\right) \perp M\left(d^*\right)|C, \tag{3.16}$$

where $Y\left(d,m\right)$ is an individual's joint potential outcome had they experienced level d of the exposure and level m of the mediator, and $M\left(d^*\right)$ is an individual's potential value of the mediator had they experienced level d^* of the exposure. Because $Y\left(d,m\right)$ and $M\left(d^*\right)$ arise from different counterfactual scenarios, they can never be observed together, no matter the data collected.

Substantively, this assumption requires that there must not be any confounders of the mediator-outcome relationship that are themselves affected by the exposure, whether these variables are observed or not. In other words, there must not be any exposure-induced confounders, as we defined them in Section 3.1. In the presence of exposure-induced confounders, the average total effect cannot be cleanly separated into a component operating through the focal mediator and a component operating through all other pathways, at least not without additional assumptions that constrain the functional form of the probability distribution from which the data were generated, and thus natural direct and indirect effects cannot be nonparametrically identified.

Assumptions (c.i) to (c.iv) would all be satisfied in data generated from a process resembling the model in Figure 3.3, where there are no unobserved causes of both D and Y, M and Y, or D and M and where there are no causes–observed or unobserved–of both M and Y that are also affected by D. In Figure 3.6, by contrast, we illustrate scenarios in which each of these assumptions would be violated. Assumption (c.i) would be violated if the data were generated from a process resembling the model in Panel A of Figure 3.6, where an unobserved variable confounds the exposure-outcome relationship. Assumption (c.ii) would be violated in Panel B, where an unobserved variable confounds the mediator-outcome relationship. In Panel C, assumption (c.iii) would be violated because there is an unobserved variable that confounds the exposure-mediator relationship. And finally, assumption (c.iv) would be violated in Panel D, where there is an exposure-induced confounder of the mediator-outcome relationship, denoted by L. In this panel, the confounder L affects the mediator M and the outcome Y, and is itself affected by the exposure D. If the data were generated from a process resembling this model, assumption (c.iv) would be violated irrespective of whether or not L is observed.

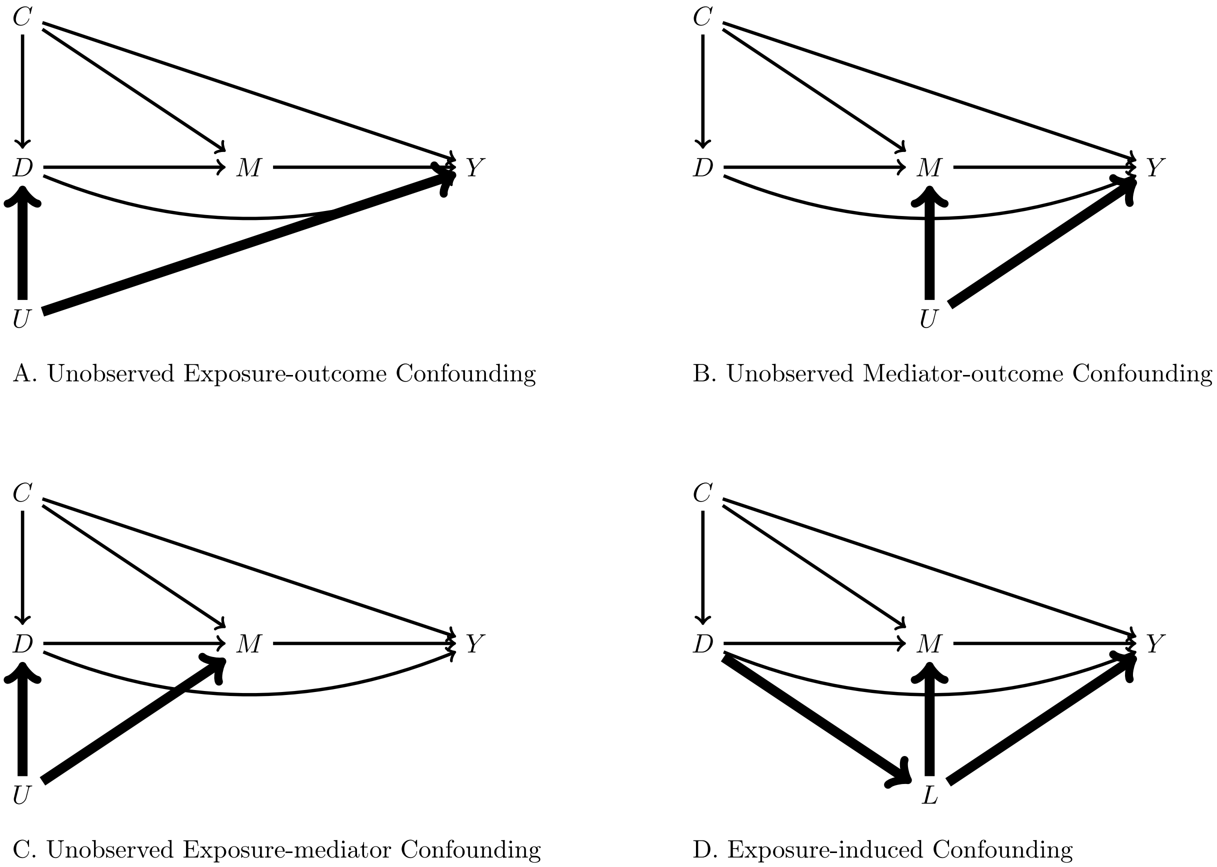
Assumption (c.v). Beyond the conditional independence assumptions outlined previously, nonparametric identification of natural direct and indirect effects also requires that there must be a positive probability of all values for the exposure and mediator conditional on the baseline confounders. This assumption is equivalent to assumption (b.iii) in Section 3.3.2 on the identifiability of controlled direct effects, and it can be formally expressed exactly as in Equation 3.12 above. Substantively, it requires that there are at least some individuals who experienced each possible level of both the exposure and mediator within every subpopulation defined by the baseline confounders.

Assumption (c.vi). Finally, nonparametric identification of the $NDE\left(d, d^{*}\right)$ and $NIE\left(d, d^{*}\right)$ depends on a multipart consistency assumption. Specifically, this assumption requires that the observed and potential values of the mediator are consistent with each other, and that the observed, potential, and nested potential outcomes are all consistent with one another. These conditions can be formally expressed as follows:

$$M = M\left(D\right) \tag{3.17}$$

$$Y = Y\left(D\right) = Y\left(D, M\left(D\right)\right), \tag{3.18}$$

where D and M denote an individual's observed values on the exposure and mediator. Equation 3.17 requires that an individual's observed mediator value M is consistent with their potential mediator value under the exposure that they did in fact experience, which is denoted by $M(D)$. Equation 3.18 requires that an individual's observed outcome Y must be equal to their potential outcome under the level of the exposure they did in fact experience, which is denoted by $Y\left(D\right)$. In addition, it requires that an individual's



A. Unobserved Exposure-outcome Confounding

B. Unobserved Mediator-outcome Confounding

C. Unobserved Exposure-mediator Confounding

D. Exposure-induced Confounding

Figure 3.6: Graphical Illustration of Unobserved and Exposure-Induced Confounding.

Note: D denotes the exposure, M denotes a mediator, Y denotes the outcome, C denotes a set of observed baseline confounders, U denotes a set of unobserved confounders, and L denotes a set of exposure-induced confounders.

potential outcome under their observed exposure, $Y(D)$, must be equal to their nested potential outcome under this same exposure and, by extension, under the value of the mediator that would arise naturally from this exposure, which is denoted by $Y(D, M(D))$.

Nonparametric identification formulas. If assumptions (c.i) to (c.vi) are all satisfied, the natural direct effect can be equated with a function of observable data rather than the nested and cross-world potential outcomes. This function, which we refer to as the nonparametric identification formula for the $NDE(d, d^*)$, can be expressed as follows:

$$NDE(d, d^*) = \sum_{m,c} \left(\mathbb{E}[Y|c, d, m] - \mathbb{E}[Y|c, d^*, m]\right) P(m|c, d^*) P(c), \tag{3.19}$$

where $\mathbb{E}[Y|c, d, m]$ denotes the conditional expected value of the observed outcome Y among individuals for whom $C = c$, $D = d$, and $M = m$, where $\mathbb{E}[Y|c, d^*, m]$ is defined analogously, and where $P(m|c, d^*)$ is the probability of experiencing level m of the mediator among those who experienced level d^* of the exposure and for whom $C = c$. As before, $P(c)$ denotes the marginal probability that $C = c$. In Appendix C, we provide a step-by-step derivation of this identification formula, beginning with the formal definition of the $NDE(d, d^*)$ in terms of the nested and cross-world potential outcomes and then invoking assumptions (c.i) to (c.vi) to arrive at the expression in Equation 3.19, which involves only the observable data.

In this expression, $\mathbb{E}[Y|c, d, m] - \mathbb{E}[Y|c, d^*, m]$ denotes the difference in the expected values of the observed outcome Y comparing individuals who experienced different levels of the exposure, $D = d$ versus $D = d^*$, but the same values on the baseline confounders and on the mediator. Within each subpopulation defined by the baseline confounders, this difference in expected values is evaluated separately among individuals with different levels of the mediator, and then these differences are averaged together, weighting each by the probability of experiencing a given level of the mediator among those who experienced level d^* of the exposure. Lastly, these weighted differences in expected values are then averaged together again, now weighting them by the probability that an individual falls in a particular subpopulation defined by the baseline confounders. In other words, the nonparametric identification formula in Equation 3.19 computes the $NDE(d, d^*)$ by perfectly stratifying the target population by all levels of the baseline confounders, evaluating the difference in the mean of the outcome between those with different levels of the exposure but the same level of the mediator, taking a weighted average of these differences in means with weights equal to the conditional probability of the mediator given exposure to d^*, and then by averaging these quantities together again across strata with weights equal to the relative size of each stratum.

Similarly, under assumptions (c.i) to (c.vi), the natural indirect effect can also be equated with a function of observable data rather than nested and cross-world potential outcomes. This function, which we refer to as the nonparametric identification formula for the $NIE(d, d^*)$, can be expressed as follows:

$$NIE(d, d^*) = \sum_{m,c} \mathbb{E}[Y|c, d, m] \left(P(m|c, d) - P(m|c, d^*)\right) P(c), \tag{3.20}$$

where $\mathbb{E}[Y|c, d, m]$, $P(m|c, d^*)$, and $P(c)$ are defined as above and where $P(m|c, d)$ is the probability of experiencing level m of the mediator among those who experienced level d of the exposure and for whom $C = c$. In Appendix D, we provide a step-by-step derivation of this identification formula, beginning with the formal definition of the $NIE(d, d^*)$ in terms of nested and cross-world potential outcomes and arriving at the expression in Equation 3.20 by invoking assumptions (c.i) to (c.vi).

In this expression, $\mathbb{E}[Y|c, d, m]$ represents the expected value of the observed outcome Y among individuals with the same values on the baseline confounders, who experienced level d of the exposure, and who experienced level m of the mediator. Within each subpopulation defined by the baseline confounders, this expected value is computed separately among individuals who experienced different levels of the mediator, and these expected values are added together, weighting each by the difference in the probability of experiencing a given level of the mediator between those with level d rather than d^* of the exposure. These weighted expected values are then averaged together again, now weighting them by the probability that an individual falls in a particular subpopulation defined by the baseline confounders. In other words, the nonparametric identification formula in Equation 3.20 computes the $NIE\left(d, d^*\right)$ by, first, perfectly stratifying the target population by all levels of the baseline confounders. Next, within each stratum, it evaluates the mean of the outcome across levels of the mediator among those with level d of the exposure. Then, it takes a weighted sum of these means, with weights equal to the difference in the probability of the mediator among those exposed to d rather than d^*. Lastly, these quantities are averaged together across strata of the confounders, with weights equal to the relative size of each stratum.

Nonparametric identification of natural direct and indirect effects hinges on a set of especially strong assumptions. Unlike the assumptions required to nonparametrically identify total and controlled direct effects, there is not, in general, any experimental design that can simultaneously satisfy all of the assumptions required to identify natural direct and indirect effects. In a conventional experiment where the exposure is randomly assigned, only assumptions (c.i) and (c.iii) would be met by design, as random assignment of individuals to different levels of an exposure ensures only that there is not any unobserved confounding of the exposure-outcome and exposure-mediator relationships. It does not ensure that the mediator-outcome relationship is unconfounded by unobserved or exposure-induced factors. It also does not ensure that all levels of both the exposure and mediator occur with positive probability, nor that the observed, potential, nested potential, and joint potential outcomes are all consistent with one another.

Moreover, even in an experiment where the exposure and the mediator are jointly or sequentially randomized, only assumptions (c.i) to (c.iii) and (c.v) would be met by design, as random assignment can ensure only that the exposure and mediator are unconfounded and that they both occur with positive probability. However, randomization of the observed exposure and mediator would not ensure that assumption (c.iv) is met, as neither $M\left(d^*\right)$ nor $Y\left(d, m\right)$ is randomized. Moreover, randomization of the mediator would violate assumption (c.vi) because $M \neq M\left(D\right)$ and thus $Y\left(D, M\left(D\right)\right) \neq Y\left(D, M\right)$ under this experimental design. In other words, when an individual's observed value on the mediator is randomly assigned, it may not equal the value that would have arisen naturally as a result of their assigned exposure, and by extension, an individual's observed outcome may not be consistent with their nested potential outcome. In Chapter 7, we elaborate on the challenges of evaluating causal mediation using randomized experiments, and we discuss alternative experimental designs that come closer to satisfying the conditions required for nonparametrically identifying natural direct and indirect effects.

Consider now the nonparametric identifiability of the $NDE\left(1, 0\right)$ and $NIE\left(1, 0\right)$ in the NLSY, which capture effects of college attendance on depression that are versus are not transmitted through a causal chain involving the risk of unemployment. Formally, both effects are nonparametrically identified if (c.i) college attendance is conditionally independent of the joint potential outcomes given the baseline confounders, (c.ii) unemployment is conditionally independent of the joint potential outcomes given the baseline confounders and college attendance, (c.iii) college attendance is conditionally independent of the potential values of the mediator given the baseline confounders, (c.iv) the joint potential outcomes under college attendance are independent of the potential values of the mediator had an individual not attended college, (c.v) there is a positive probability of college attendance and unemployment within each of the subpopulations defined by the baseline confounders, and (c.vi) the observed, potential, nested potential, and joint potential outcomes are consistent with one another.

In substantive terms, assumptions (c.i) to (c.iii) require that there not be any unobserved factors that confound the relationships between college attendance and depression, unemployment and depression, or college attendance and unemployment, while assumption (c.iv) requires that there must not be any factors, whether observed or not, that confound the relationship of unemployment with depression and that are also affected by college attendance. Assumption (c.v) additionally requires that, within every subgroup defined by the baseline confounders, there are at least some individuals who experienced each combination of educational attainment and employment status in the target population. Finally, assumption (c.vi) requires that there are not multiple versions of the exposure or mediator with heterogeneous effects on the outcome, and that there is no interference between individuals.

These are exacting assumptions. They are unlikely to hold in our analysis of the NLSY because there are many unobserved variables that may jointly affect college attendance, unemployment, and depression and also because "college attendance" and "unemployment" both involve diverse experiences with potentially variable effects on mental health that spillover from one individual to another. Indeed, these assumptions may not hold in any observational study, and an experimental design capable of satisfying them all simultaneously does not exist. These challenges underscore the difficulty of analyzing causal mediation no matter the research design, and they warrant at least some measure of skepticism about all analyses of natural direct and indirect effects. In this situation, researchers must take care to assess the sensitivity of their inferences to potential violations of these assumptions, as we discuss in Section 3.7 below.

3.4 Nonparametric Estimation

We have thus far reduced causal mediation analysis to an identification problem. In the previous section, our goal was to equate causal effects defined in terms of counterfactuals with empirical quantities defined in terms of observable data, all while ignoring any random variability that may arise from sampling a target population instead of observing all its members. In practice, however, we rarely have data from the entire target population and thus cannot simply ignore random variability. Rather, in most applications of causal mediation analysis, researchers only have data from a random sample of the target population. In this section and henceforth, we shift our focus from identification with full population data to estimation with data from a random sample.

In general terms, estimation involves using sample data from the target population to learn about the causal effects of interest. More specifically, a *point estimator* is a function that takes data from a sample as input and then outputs a single numeric value for the focal estimand. A *point estimate*, by extension, is the particular numeric value given by an estimator when applied to data from one specific sample. Because an estimator is a function of a random sample, it is a random variable, and it has a probability distribution, expected value, variance, and so on, just like any other random variable. The probability distribution of an estimator is known as the *sampling distribution.* This distribution describes all the possible estimates and their associated probabilities if an estimator were applied, repeatedly, to data from all possible random samples of size n from a target population of size N. In other words, if we were to draw all possible random samples of a given size from the target population and then apply the estimator to each sample and obtain an estimate, the probability distribution of all these different estimates is the sampling distribution.

Any given estimate may differ from the target estimand because it may suffer from two different sources of error: bias and sampling variability. *Bias* is a type of systemic error that arises when the assumptions required to identify an estimand are violated. Formally, an estimator, denoted by $\hat{\theta}$, for a target estimand, denoted by θ, is biased if $\mathbb{E}\left[\hat{\theta}\right] \neq \theta$, and it is unbiased if $\mathbb{E}\left[\hat{\theta}\right] = \theta$. In substantive terms, if the expected value, or average, of an estimator is equal to its estimand, then it is unbiased; otherwise, it is biased and will yield estimates that systematically differ from the target estimand. With an unbiased estimator, the sampling distribution of its estimates is centered around the target estimand.

Sampling variability is a source of random rather than systemic error. Even if an estimator is unbiased, any given estimate that it yields when applied to a particular random sample may differ from the target estimand because the sample does not include all members of the target population. The discrepancy between an estimand and an estimate given by an unbiased estimator is known as a sampling error.

The magnitude of sampling errors with an unbiased estimator can be summarized by its variance–that is, by the dispersion of its sampling distribution. Unbiased estimators that tend to yield smaller sampling errors, and thus have lower variance, are said to be *efficient*. Specifically, if two different estimators are both unbiased and applied to samples of the same size, one estimator $\hat{\theta}$ is said to be relatively more efficient than another estimator $\tilde{\theta}$ for the same target estimand when $Var\left[\hat{\theta}\right] \leq Var\left[\tilde{\theta}\right]$ for all values of θ, where $Var\left[\cdot\right]$ denotes the variance and where strict inequality holds for at least one value of θ. In substantive terms, if one estimator is more efficient than another, the sampling distribution of its estimates is less spread out.

Relatedly, an estimator is said to be *consistent* if the estimates it yields get arbitrarily close to the target estimand as the sample size increases. Formally, an estimator $\hat{\theta}$ is consistent if it converges in probability to the true value of its target estimand: $\underset{n\rightarrow\infty}{lim} P\left(\left|\hat{\theta}-\theta\right| > \epsilon\right) = 0$ for any $\epsilon > 0$. In substantive terms, with a consistent estimator, the difference between the estimates it yields when applied to data from a random sample and the target estimand will converge to zero as the sample size becomes ever larger. Or in other words, as the number of observations increases, the sampling distribution collapses on the true value of the target estimand. Note that, despite their similar terminology, the consistency property of an estimator is fundamentally different from a consistency assumption about the observed and potential outcomes.

In this section, we explain how to construct consistent estimates for total, direct and indirect effects without invoking any additional assumptions beyond those required to identify these effects nonparametrically. That is, we explain how to estimate causal effects without using any parametric models that might explicitly or implicitly impose functional form restrictions on the probability distribution from which the observed data are sampled.

We illustrate this approach using an artificially simplified example, as nonparametric estimation of total, direct, and indirect effects is often impossible or impractical due to sparsity in the sample data, the curse of dimensionality, or other related complexities. Moreover, even when nonparametric estimation is possible with the available sample data, there are still reasons why an analyst might prefer an estimation strategy that relies on parametric models, such as greater relative efficiency, despite the additional assumptions required. We revisit these issues to conclude the section. Nevertheless, we illustrate nonparametric estimation in detail because, when it is feasible and practical to implement, this approach can yield consistent estimates under weaker assumptions than parametric approaches.

Essentially, nonparametric estimation just involves plugging in sample analogs (i.e., sample means and proportions) for the population quantities (i.e., expected values and probabilities) in the identification formulas from Section 3.3 above. Suppose, for the purpose of illustration, that we had only a single binary confounder in the NLSY–whether an individual's mother ever attended college–for which adjustment is re-

Table 3.1: Case Counts and Sample Means for CES-D Scores (Y) by College Attendance (D), Unemployment Status (M), and Maternal Education (C), NLSY.

		Maternal Education			
Respondent Education	Unemployment Status	No College ($C = 0$)		Attended College ($C = 1$)	
		$\bar{Y}_{c,d,m}$	$n_{c,d,m}$	$\bar{Y}_{c,d,m}$	$n_{c,d,m}$
No College	Never Unemployed ($M = 0$)	.00	1836	−.02	178
($D = 0$)	Ever Unemployed ($M = 1$)	.32	548	.37	40
	Total ($\bar{Y}_{c,d}$; $n_{c,d}$)	.07	2384	.05	218
Attended College	Never Unemployed ($M = 0$)	−.22	594	−.17	347
($D = 1$)	Ever Unemployed ($M = 1$)	.08	96	−.07	43
	Total ($\bar{Y}_{c,d}$; $n_{c,d}$)	−.18	690	−.16	390

Note: CES-D scores have been standardized to have zero mean and unit variance. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/table_3-1`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

quired to identify the total, direct, and indirect effects of interest. In this situation, all the information we need to construct nonparametric estimates is contained in Table 3.1, which reports case counts and sample means for the CES-D, separately by maternal education, respondent education, and unemployment status, for $n = 3682$ individuals in the NLSY.

In this table, $\bar{Y}_{c,d,m}$ denotes the sample mean of standardized CES-D scores among NLSY respondents for whom $C = c$, $D = d$, and $M = m$, where these variables respectively denote whether a sample member's mother attended college, whether the sample member attended college themselves, and whether the sample member experienced a spell of unemployment. Similarly, $n_{c,d,m}$ just denotes the number of respondents for whom $C = c$, $D = d$, and $M = m$ in the NLSY. Thus, for example, there are $n_{1,0,0} = 178$ respondents who did not attend college or experience a spell of unemployment but who had mothers that did attend college, and their average CES-D score is $\bar{Y}_{1,0,0} = -.02$ standard deviations–that is, just below the total sample mean. The quantities denoted by $\bar{Y}_{c,d}$ and $n_{c,d}$ also represent sample means and counts but now among respondents for whom $C = c$ and $D = d$, collapsing over values of the mediator.

With these data, a nonparametric estimator for the average total effect can be expressed as follows:

$$\begin{aligned}\widehat{ATE}\left(d, d^*\right)^{np} &= \sum_c \left(\hat{\mathbb{E}}\left[Y|c, d\right] - \hat{\mathbb{E}}\left[Y|c, d^*\right]\right) \hat{P}\left(c\right) \\ &= \sum_c \left(\bar{Y}_{c,d} - \bar{Y}_{c,d^*}\right) \hat{\pi}_c, \qquad (3.21)\end{aligned}$$

where $\hat{\pi}_c = \sum_d n_{c,d}/n$ denotes the proportion of sample members for whom $C = c$. In this expression and henceforth, "hats" distinguish between an estimator and its estimand, while superscripts distinguish between different estimators for the same estimand. The "np" superscript in Equation 3.21 stands for "nonparametric."

Using this expression, a nonparametric estimate for the average total effect of college attendance on depression can be obtained from the data in Table 3.1 as follows:

$$
\begin{aligned}
\widehat{ATE}\left(1,0\right)^{np} &= \left(\bar{Y}_{0,1} - \bar{Y}_{0,0}\right)\hat{\pi}_0 + \left(\bar{Y}_{1,1} - \bar{Y}_{1,0}\right)\hat{\pi}_1 \\
&= (-.18 - .07)\left(\frac{2384 + 690}{3682}\right) + (-.16 - .05)\left(\frac{218 + 390}{3682}\right) \\
&= -.24,
\end{aligned}
$$

which suggests that attending college reduces CES-D scores by about 0.24 standard deviations, on average. This estimate is a weighted sum of differences in mean CES-D scores comparing sample members who did attend college with those who did not. These differences are evaluated separately among sample members with mothers who did not attend college and among those with mothers who did, and then they are combined using weights equal to the proportion of sample members with mothers at each level of education. In other words, to compute $\widehat{ATE}\left(d, d^*\right)^{np}$, we stratify the sample by C, take the difference in the mean of the outcome between those with different values of the exposure, and then sum the stratum-specific differences in means, weighting each by the relative size of each stratum.

Similarly, a nonparametric estimator for controlled direct effects can be expressed as follows:

$$
\begin{aligned}
\widehat{CDE}\left(d, d^*, m\right)^{np} &= \sum_c \left(\hat{\mathbb{E}}\left[Y|c, d, m\right] - \hat{\mathbb{E}}\left[Y|c, d^*, m\right]\right)\hat{P}\left(c\right) \\
&= \sum_c \left(\bar{Y}_{c,d,m} - \bar{Y}_{c,d^*,m}\right)\hat{\pi}_c, \qquad (3.22)
\end{aligned}
$$

where $\bar{Y}_{c,d,m}$ and $\hat{\pi}_c$ are defined as above. Using this expression, an estimate for the controlled direct effect of college attendance on depression, if individuals were never to experience a spell of unemployment, can be constructed from Table 3.1 as follows:

$$
\begin{aligned}
\widehat{CDE}\left(1,0,0\right)^{np} &= \left(\bar{Y}_{0,1,0} - \bar{Y}_{0,0,0}\right)\hat{\pi}_0 + \left(\bar{Y}_{1,1,0} - \bar{Y}_{1,0,0}\right)\hat{\pi}_1. \\
&= (-.22 - .00)\left(\frac{2384 + 690}{3682}\right) + (-.17 - (-.02))\left(\frac{218 + 390}{3682}\right) \\
&= -.21,
\end{aligned}
$$

which suggests that attending college would reduce CES-D scores by about 0.21 standard deviations, on average, if individuals avoided unemployment.

This estimate is a weighted sum of differences in mean CES-D scores comparing sample members who attended college with those who did not attend college among the subset of participants who never experienced unemployment. These differences are evaluated separately among sample members with mothers who did not attend college and among those with mothers who did, and then they are combined using weights equal to the proportion of sample members with mothers at each level of education. In other words, to compute $\widehat{CDE}\left(d, d^*, m\right)^{np}$, we stratify the sample by C, take the difference in the mean of the outcome between those with different values of the exposure but the same level of the mediator, and then sum the stratum-specific differences in means, weighting each by the relative size of each stratum.

A nonparametric estimator for the natural direct effect can be expressed as follows:

$$\begin{aligned}\widehat{NDE}\left(d,d^{*}\right)^{np} &= \sum_{m,c}\left(\hat{\mathbb{E}}\left[Y|c,d,m\right]-\hat{\mathbb{E}}\left[Y|c,d^{*},m\right]\right)\hat{P}\left(m|c,d^{*}\right)\hat{P}\left(c\right)\\ &= \sum_{m,c}\left(\bar{Y}_{c,d,m}-\bar{Y}_{c,d^{*},m}\right)\hat{\pi}_{m|c,d^{*}}\hat{\pi}_{c},\end{aligned} \tag{3.23}$$

where $\hat{\pi}_{m|c,d^{*}} = {}^{n_{c,d^{*},m}}/_{n_{c,d^{*}}}$ is the proportion of sample members for whom $M = m$ among those with $C = c$ and $D = d^{*}$. With this expression, an estimate for the natural direct effect of college attendance on depression can be constructed from Table 3.1 as follows:

$$\begin{aligned}\widehat{NDE}\left(1,0\right)^{np} &= \left(\left(\bar{Y}_{0,1,0}-\bar{Y}_{0,0,0}\right)\hat{\pi}_{0|0,0}+\left(\bar{Y}_{0,1,1}-\bar{Y}_{0,0,1}\right)\hat{\pi}_{1|0,0}\right)\hat{\pi}_{0}\\ &\quad+\left(\left(\bar{Y}_{1,1,0}-\bar{Y}_{1,0,0}\right)\hat{\pi}_{0|1,0}+\left(\bar{Y}_{1,1,1}-\bar{Y}_{1,0,1}\right)\hat{\pi}_{1|1,0}\right)\hat{\pi}_{1}\\ &= \left((-.22-.00)\left(\frac{1836}{2384}\right)+(.08-.32)\left(\frac{548}{2384}\right)\right)\left(\frac{2384+690}{3682}\right)\\ &\quad+\left((-.17-(-.02))\left(\frac{178}{218}\right)+(-.07-.37)\left(\frac{40}{218}\right)\right)\left(\frac{218+390}{3682}\right)\\ &= -.22,\end{aligned}$$

which suggests that attending college would reduce CES-D scores by 0.22 standard deviations, on average, even if individuals experienced the level of unemployment that would have occurred for them had they not attended college.

To compute this estimate, we first stratify the sample by maternal education and by respondent unemployment status. Next, within levels of both these variables, we take the difference in mean CES-D scores comparing sample members who attended college with those who did not. Then, within levels of maternal education, we sum these differences across levels of the mediator, weighting them by the sample distribution of unemployment among those who did not attend college. Finally, these differences are combined again, now across levels of maternal education and with weights given by the proportion of sample members with mothers at each level of education.

In addition, a nonparametric estimator for the natural direct effect is given by the following expression:

$$\begin{aligned}\widehat{NIE}\left(d,d^{*}\right)^{np} &= \sum_{m,c}\hat{\mathbb{E}}\left[Y|c,d,m\right]\left(\hat{P}\left(m|c,d\right)-\hat{P}\left(m|c,d^{*}\right)\right)\hat{P}\left(c\right)\\ &= \sum_{m,c}\bar{Y}_{c,d,m}\left(\hat{\pi}_{m|c,d}-\hat{\pi}_{m|c,d^{*}}\right)\hat{\pi}_{c}.\end{aligned} \tag{3.24}$$

In Equation 3.24, $\hat{\pi}_{m|c,d^{*}} = {}^{n_{c,d^{*},m}}/_{n_{c,d^{*}}}$ is the proportion of sample members for whom $M = m$ among those with $C = c$ and $D = d^{*}$, as above; $\hat{\pi}_{m|c,d}$ is defined analogously, and thus $\hat{\pi}_{m|c,d} - \hat{\pi}_{m|c,d^{*}}$ is the difference in the proportion of sample members for whom $M = m$ comparing those exposed to d rather than d^{*} within level c of the baseline confounder. Using this expression, a nonparametric estimate for the natural indirect effect of college attendance on depression can be constructed from Table 3.1 as follows:

$$\begin{aligned}\widehat{NIE}\left(1,0\right)^{np} &= \left(\bar{Y}_{0,1,0}\left(\hat{\pi}_{0|0,1}-\hat{\pi}_{0|0,0}\right)+\bar{Y}_{0,1,1}\left(\hat{\pi}_{1|0,1}-\hat{\pi}_{1|0,0}\right)\right)\hat{\pi}_{0} \\ &\quad +\left(\bar{Y}_{1,1,0}\left(\hat{\pi}_{0|1,1}-\hat{\pi}_{0|1,0}\right)+\bar{Y}_{1,1,1}\left(\hat{\pi}_{1|1,1}-\hat{\pi}_{1|1,0}\right)\right)\hat{\pi}_{1} \\ &= \left(-.22\left(\frac{594}{690}-\frac{1836}{2384}\right)+.08\left(\frac{96}{690}-\frac{548}{2384}\right)\right)\left(\frac{2384+690}{3682}\right) \\ &\quad +\left(-.17\left(\frac{347}{390}-\frac{178}{218}\right)-.07\left(\frac{43}{390}-\frac{40}{218}\right)\right)\left(\frac{218+390}{3682}\right) \\ &= -0.02,\end{aligned}$$

which suggests that if individuals experienced the level of unemployment that would have occurred for them had they attended college, rather than the level of unemployment that would have occurred had they not attended college, their CES-D scores would only decline by 0.02 standard deviations.

To compute this estimate, we first stratify the sample by maternal education. Next, within levels of maternal education, we evaluate the mean CES-D score at each level of unemployment among sample members who attended college. Then, we sum these means across levels of the mediator, weighting by the difference in the sample distribution of unemployment among respondents with versus without a college education. Finally, we combine these quantities across levels of maternal education using weights equal to the proportion of sample members with mothers at each level of education.

The nonparametric estimates for the natural direct and indirect effects sum to the nonparametric estimate for the average total effect, as expected. Based on the data from Table 3.1 $\widehat{ATE}\left(1,0\right)^{np}=\widehat{NDE}\left(1,0\right)^{np}+\widehat{NIE}\left(1,0\right)^{np}=-.22-.02=-.24$. Taken altogether, these estimates suggest that college attendance reduces depression at midlife overall, but relatively little of this effect appears to arise from mediation via differences in the risk of unemployment.

If the total, direct, and indirect effects of interest are nonparametrically identified, then their corresponding nonparametric estimators are consistent–that is, as the size of the sample to which they are applied becomes large, these estimators get closer, and closer, and ultimately converge to their target estimands. Nonparametric identification, and by extension, the consistency property of these estimators depends on a number of strong assumptions about the absence of unobserved and exposure-induced confounding. In our artificially simplified illustration, these assumptions strain credulity: we adjusted only for a single binary confounder–whether an individual's mother attended college– when it is likely that many other factors jointly affect an individual's education, unemployment status, and risk of depression. If the assumptions that motivate nonparametric estimation are violated, the estimators described above may be biased and inconsistent for their target estimands, or in other words, they may not suffer merely from random error due to sampling variability but also from systematic error owing to confounding bias. As a result, any inferences that we might draw about the extent or nature of causal mediation would likely be mistaken.

But even if there is no unobserved or exposure-induced confounding, problems with nonparametric estimation may still arise. We implemented the nonparametric estimators using an artificially simplified example mainly for illustrative purposes but also to highlight difficulties that frequently arise with this approach in practice. These difficulties include sparsity, the curse of dimensionality, and high sampling variability, which all stem from the use of finite sample data containing many confounders and/or variables with many values.

Nonparametric estimation involves perfectly stratifying the sample data by all levels of the baseline confounders. In our empirical illustration, this merely involved dividing the sample into those with and without mothers who attended college. But suppose that there were many confounders or that each confounder had many levels, as is typical of social science applications. When there are many levels or combinations of different confounders, there may be few strata that contain individuals who experience all levels of the exposure and mediator, that is, there may be random departures from the assumption of positivity. In fact, when there are a very large number of confounders or these variables are truly continuous, perfectly stratifying the sample data can produce as many strata as there are observations, leaving no comparison cases to estimate the effects of interest within each stratum. In this situation, nonparametric estimation could not be performed due to the high dimensionality of the baseline confounders and the data sparsity it engenders.

Nonparametric estimation also involves comparing outcome means across levels of the exposure and mediator. In our empirical illustration, where the exposure and mediator were both binary, this merely involved computing and comparing four different outcome means within each stratum of the baseline confounder–one for every combination of respondent education and unemployment status. But now suppose that the exposure or mediator had many values. When there are many levels or combinations of the exposure and mediator, there may be values for these variables that are not actually experienced by any individual in the sample, which would preclude computing the corresponding outcome mean. Moreover, when the exposure or mediator are truly continuous, then the sample mean of the outcome may be undefined for nearly all values of these variables because they are infinitely divisible. In this situation, nonparametric estimation also could not be performed due to sparsity in the sample data, and relatedly, to departures from the assumption of positivity.

Finally, suppose that the sample data include individuals who experienced all levels of both the exposure and mediator within every stratum of the baseline confounders, as in our simplified illustration where nonparametric estimation was possible. But now suppose further that there were only a small number of sample members within each subgroup defined by the exposure, mediator, and confounders. In this situation, nonparametric estimates for the total, direct, and indirect effects of interest may suffer from a high degree of sampling variability because the terms that compose them are each computed using a small number of observations. Thus, even when nonparametric estimation is feasible, it can still yield highly imprecise estimates in finite samples.

To conclude, when nonparametric estimation is feasible and based on a sufficiently large sample with a low-dimensional set of variables, this approach can yield consistent and relatively precise estimates of total, direct, and indirect effects under assumptions that are weaker than those required of all the other methods we consider throughout the rest of the chapter. However, in most social science applications, nonparametric estimation is either impossible or impractical due to data sparsity, high dimensionality, or excessive sampling variability. When these problems complicate or preclude a nonparametric approach to estimation, an alternative approach that relies on parametric models may offer considerable advantages, despite the stronger assumptions required. We introduce a set of parametric estimators and revisit these trade-offs in Section 3.5 below.

3.5 Parametric Estimation

In this section, we explain how to construct consistent estimates for total, direct, and indirect effects using parametric models. A *parametric model* is defined by the restrictions that it imposes on the joint distribution of the observed data. For example, consider the familiar linear regression model for the conditional expected value of an outcome Y given the exposure D and baseline confounders C, which can be expressed as follows:

$$\mathbb{E}\left[Y|c,d\right] = \alpha_0 + \alpha_1^T c + \alpha_2 d, \tag{3.25}$$

where α_0 is known as the intercept or constant term, α_1^T is a transposed vector of coefficients multiplying each of the baseline confounders, and α_2 is the coefficient on the exposure. Note that there is no error term in this equation because it is expressed as a model for the conditional expected value, $\mathbb{E}\left[Y|c,d\right]$, rather than a model for the individual outcomes. An error term is only relevant when modeling the individual outcomes, which are additionally subject to random variation around their conditional expectation.

The model in Equation 3.25 encodes a restriction on the shape of the conditional expected value $\mathbb{E}\left[Y|c,d\right]$, and by extension, on the joint distribution of the data. Specifically, it restricts the conditional expected value to be a linear and additive function, such that the many different values of $\mathbb{E}\left[Y|c,d\right]$ are now given by combinations of a relatively small number of parameters $\left\{\alpha_0, \alpha_1^T, \alpha_2\right\}$. Through these types of restrictions, parametric models can compensate for data limitations that complicate or preclude the use of nonparametric methods.

A *parametric estimator* is based on a parametric model or a set of parametric models. A model like Equation 3.25 is typically estimated by the method of ordinary least squares (OLS). The OLS estimator selects estimates for model parameters by finding values that minimize the sum of squared prediction errors in the sample data. Because OLS uses data from all sample members to find estimates for these parameters, our fitted model can be used to estimate quantities that we might not be able to compute with nonparametric methods.

To appreciate this, suppose that there are several strata of the baselines confounders in which all sample members experienced the same exposure d and none experienced the alternative exposure d^*. In this situation, nonparametric estimation of every value for $\mathbb{E}\left[Y|c,d^*\right]$ would not be possible, as there are strata of the confounders without any observations for whom $D = d^*$ in the sample. Nevertheless, using a model like Equation 3.25 fit by OLS, we could still obtain an estimate of $\mathbb{E}\left[Y|c,d^*\right]$ for every stratum of the confounders, even for those without any sample members who were actually exposed to d^*. This is because our model assumes that all the conditional expected values are given by a linear function of just a few parameters, and OLS leverages all the sample data to estimate them. Essentially, OLS estimation of these conditional means by way of a parametric model involves borrowing information from sample members who were exposed to d^* in other strata to fill in information that is missing from the sample elsewhere. In general, choices among parametric estimators and the different models on which they are based determine how information is shared across the data space.

The ability of parametric estimators to overcome the problem of sparsity by sharing, borrowing, and filling in information that is otherwise not available from the sample data alone comes at a cost, however. In particular, parametric estimators are only consistent if their underlying models are correctly specified–that is, if the restrictions these models impose on the distribution of the observed data are accurate. If the models on which parametric estimators rely are misspecified, then they will suffer from systematic bias, and they will not converge to their target estimands as the sample size becomes large.

For example, suppose that Equation 3.25 were misspecified, and that the correct model for the expected value of the outcome conditional on the exposure and confounders is instead given by the following expression:

$$\mathbb{E}\left[Y|c,d\right] = \alpha_0 + \alpha_1^T c + d\left(\alpha_2 + \alpha_3^T c\right), \tag{3.26}$$

where α_3^T is a transposed vector of coefficients multiplying all the two-way interactions between the exposure and each of the baseline confounders. In this situation, any estimates based on the simpler model in Equation 3.25 will suffer from systematic bias.

With parametric estimation, then, the positivity conditions required of nonparametric estimation are supplanted by an alternative set of assumptions about correct model specification. Because parametric models are rarely, if ever, correctly specified, some degree of misspecification is almost always expected. In this situation, the practical goal when implementing a parametric estimator is to find a model or set of models that are approximately correct and are therefore less likely to yield estimates that suffer from severe bias.

Achieving this goal, however, requires navigating a trade-off between bias and variance. In general, the larger the number of parameters in a model, the fewer restrictions it imposes on the joint distribution of the observed data, and the less it borrows information from one location in the data space to fill in information missing from other areas. Estimators based on complex models with many parameters therefore afford greater protection against bias due to model misspecification. Nevertheless, complex models with many parameters also tend to produce estimates that suffer from greater sampling variability. Thus, when analyzing causal mediation, researchers must balance the desire for protection against bias that is afforded by more complex models with concerns about imprecision due to high variance.

Throughout the remainder of this section, we outline a set of parametric estimators for total, direct, and indirect effects. We begin with estimators based exclusively on linear models, followed by an approach based on a broader class of both linear and nonlinear models for the mediator and outcome. These approaches are flexible and can be used with either discrete or continuous exposures, as well as various types of mediators and outcomes. We then present another set of parametric estimators based on inverse probability weights, which are constructed from models for the exposure. This approach is best suited for applications where the exposure is binary or has a relatively small number of discrete values.

With all these estimators, whether they are consistent for the effects of interest hinges not only on the identification assumptions discussed previously but also on the absence of any model misspecification. In Chapter 6, we describe a class of robust methods for estimating total, direct, and indirect effects that mitigate the problem of model misspecification by constructing models inductively using data-adaptive machine learning algorithms.

3.5.1 Estimation with Linear Models

In this section, we explain how to estimate total, direct, and indirect effects with linear models for both the mediator and the outcome. Consider first the following set of linear models:

$$\mathbb{E}\left[M|c,d\right] = \beta_0 + \beta_1^T c + \beta_2 d \tag{3.27}$$

$$\mathbb{E}\left[Y|c,d,m\right] = \gamma_0 + \gamma_1^T c + \gamma_2 d + \gamma_3 m, \tag{3.28}$$

where $\mathbb{E}\left[M|c,d\right]$ denotes the conditional expected value of the mediator given the baseline confounders and the exposure and $\mathbb{E}\left[Y|c,d,m\right]$ denotes the conditional expected value of the outcome given the confounders, exposure, and mediator. If assumptions (c.i) to (c.vi) are satisfied and Equations 3.27 and 3.28 are both correctly specified, then the total, direct, and indirect effects of interest are given by simple functions of the parameters in these models. In particular, the natural direct and indirect effects are given by $NDE\left(d,d^*\right) =$

$\gamma_2 (d - d^*)$ and $NIE(d, d^*) = \beta_2\gamma_3 (d - d^*)$, respectively, and the average total effect is given by their sum, such that $ATE(d, d^*) = (\gamma_2 + \beta_2\gamma_3)(d - d^*)$. Moreover, because there is no interaction between the exposure and mediator in the model for the outcome, the controlled direct effect is equal to the natural direct effect, where $CDE(d, d^*, m) = \gamma_2 (d - d^*)$ as well.

These expressions may look familiar. They were the workhorse of mediation analysis for decades (Alwin and Hauser 1975; Baron and Kenny 1986; Duncan 1966; Goldberger and Duncan 1973; Sobel 1982), and they still form the backbone of several modern texts on the subject (Hayes 2017; MacKinnon 2008). Equation 3.27 is just a conventional linear regression of the mediator on the exposure and the baseline confounders, while Equation 3.28 is just a conventional linear regression of the outcome on the confounders, exposure, and mediator. Given these models, the natural and controlled direct effects are equivalent, where both are equal to the coefficient on the exposure in the regression for the outcome, γ_2, multiplied by the exposure contrast of interest, $d - d^*$. The natural indirect effect is just the product of the coefficient on the exposure in the regression for the mediator, β_2, and the coefficient on the mediator in the regression for the outcome, γ_3, multiplied by the exposure contrast of interest, $d - d^*$.

Where do these parametric expressions come from? In many analyses of mediation, they appear out of thin air, without reference to a target estimand defined using potential outcomes notation and without any discussion of identification conditions. In fact, the parametric expressions outlined previously for the total, direct, and indirect effects of interest just come from appropriately substituting linear and additive models for the mediator and outcome into the nonparametric identification formulas for these estimands. In Appendix E, we provide a step-by-step derivation of these parametric expressions for illustrative purposes, beginning with a set of nonparametric identification formulas, substituting Equations 3.27 and 3.28 into them, and then simplifying to arrive at the concise functions of model coefficients given above.

Under the identification assumptions outlined in Section 3.3 and under the additional assumption of correct model specification, consistent estimators for total, direct, and indirect effects can be constructed by fitting Equations 3.27 and 3.28 using the method of OLS and then by substituting their coefficient estimates into the appropriate parametric expressions. Specifically, a set of consistent estimators can be constructed as follows:

$$\begin{aligned}
\widehat{NDE}(d, d^*)^{lm} &= \widehat{CDE}(d, d^*, m)^{lm} = \hat{\gamma}_2 (d - d^*) \\
\widehat{NIE}(d, d^*)^{lm} &= \hat{\beta}_2\hat{\gamma}_3 (d - d^*) \\
\widehat{ATE}(d, d^*)^{lm} &= \left(\hat{\gamma}_2 + \hat{\beta}_2\hat{\gamma}_3\right)(d - d^*), \qquad (3.29)
\end{aligned}$$

where the "hats" distinguish estimators from estimands, as before, and where the "lm" superscript indicates that these estimators are based on linear and additive models fit to the sample data by least squares.

Although the parametric estimators in Equation 3.29 have been widely used in analyses of causal mediation, they are based on highly restrictive models, and thus they are prone to misspecification bias. In particular, the models on which these estimators are based do not allow the effects of the exposure and mediator to interact or to vary across levels of the baseline confounders. If the exposure and mediator interact or their effects are moderated by the confounders, then the estimators based on Equations 3.27 and 3.28 will be biased and inconsistent, as the relationship of these variables with each other and with the outcome is constrained to be additive, when in fact it is not.

Fortunately, estimation with linear models can be easily adapted to accommodate interaction effects between the exposure and mediator. Consider next the following set of linear models for the mediator and outcome:

$$\mathbb{E}[M|c,d] = \beta_0 + \beta_1^T c^{\perp} + \beta_2 d \tag{3.30}$$

$$\mathbb{E}[Y|c,d,m] = \gamma_0 + \gamma_1^T c^{\perp} + \gamma_2 d + m(\gamma_3 + \gamma_4 d). \tag{3.31}$$

These models are nearly identical to Equations 3.27 and 3.28 above except for two important differences. First, in the outcome model, we have now included an exposure-mediator interaction effect given by γ_4. Second, in both models, we have centered the baseline confounders around their sample means, where $c^{\perp} = c - \bar{C}$.

If assumptions (c.i) to (c.vi) are satisfied and Equations 3.30 and 3.31 are both correctly specified, then the natural direct and indirect effects are now given by $NDE(d,d^*) = (\gamma_2 + \gamma_4(\beta_0 + \beta_2 d^*))(d - d^*)$ and $NIE(d,d^*) = \beta_2(\gamma_3 + \gamma_4 d)(d - d^*)$, respectively, while the total effect is given by the sum of these two expressions. Moreover, because the outcome model incorporates an interaction between the exposure and mediator, the parametric expression for the controlled direct effect now differs from that for the natural direct effect. In particular, the controlled direct effect is given by $CDE(d,d^*,m) = (\gamma_2 + \gamma_4 m)(d - d^*)$ under this specification for the outcome model. As before, these parametric expressions just come from substituting the models in Equations 3.30 and 3.31 into the nonparametric identification formulas and then appropriately simplifying them.

The models in Equations 3.30 and 3.31 are more flexible in that they allow for an exposure-mediator interaction effect on the outcome, but they still constrain the effects of the exposure and mediator to be invariant across levels of the baseline confounders. In many social science applications, where effect heterogeneity is endemic, this constraint may also be unreasonable, leading to misspecification bias.

Fortunately, estimation with linear models can also be adapted to accommodate effect moderation by the baseline confounders. Consider now the following set of models, which allow the effects of the exposure and mediator to vary across levels of the confounders:

$$\mathbb{E}[M|c,d] = \beta_0 + \beta_1^T c^{\perp} + d\left(\beta_2 + \beta_3^T c^{\perp}\right) \tag{3.32}$$

$$\mathbb{E}[Y|c,d,m] = \gamma_0 + \gamma_1^T c^{\perp} + \gamma_2 d + m(\gamma_3 + \gamma_4 d) + c^{\perp}\left(\gamma_5^T d + m\left(\gamma_6^T + \gamma_7^T d\right)\right). \tag{3.33}$$

These models are similar to those in Equations 3.30 and 3.31 above except for a few important differences. Specifically, in the model for the mediator, we have additionally incorporated interactions between d and $c^{\perp}$, where β_3^T is a transposed vector of coefficients multiplying all two-way interactions of the exposure with each of the mean-centered confounders. And in the model for the outcome, we have additionally incorporated interactions of d and m with $c^{\perp}$, where the coefficient vectors denoted by $\left\{\gamma_5^T, \gamma_6^T, \gamma_7^T\right\}$ allow the joint effects of both the exposure and mediator to differ across levels of the baseline confounders.

Provided that all of these new interaction terms are constructed after mean-centering the baseline confounders, and not with their original values, the parametric expressions for total, direct, and indirect effects based on Equations 3.32 and 3.33 are exactly the same as those provided above based on Equations 3.30 and 3.31. That is, the parametric expressions for the effects of interest are unchanged by incorporating confounder-by-exposure and confounder-by-mediator interactions in our linear models, as long as these interactions are constructed with the mean-centered confounders (Wodtke et al. 2020; Wodtke and Zhou 2020). Adding these terms merely relaxes modeling restrictions on the joint distribution of the observed data by

Table 3.2: Total, Direct, and Indirect Effects of College Attendance on CES-D Scores as Estimated from Linear Models Fit to the NLSY.

	Point Estimates		
Estimand	Additive Linear Model (LinMod)	LinMod with $D \times M$ Interaction Term	LinMod with $D \times M$, $C \times D$ and $C \times M$ Interaction Terms
$ATE(1,0)$	$-.079$	$-.083$	$-.129$
$NDE(1,0)$	$-.072$	$-.078$	$-.119$
$NIE(1,0)$	$-.007$	$-.005$	$-.010$
$CDE(1,0,0)$	$-.072$	$-.054$	$-.111$

Note: Estimates are expressed in standard deviation units. D denotes college attendance, M unemployment status, and C the baseline confounders. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/table_3-2`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

allowing for certain patterns of effect moderation across levels of the baseline confounders, potentially mitigating bias due to misspecification.

Under models resembling those in Equations 3.30 to 3.33, consistent estimators for the effects of interest can be constructed by fitting these models to sample data using OLS and then by substituting their coefficient estimates into the parametric expressions provided above. Specifically, a set of consistent estimators can be constructed from the fitted models as follows:

$$\begin{aligned}
\widehat{NDE}\left(d,d^*\right)^{lmi} &= \left(\hat{\gamma}_2 + \hat{\gamma}_4\left(\hat{\beta}_0 + \hat{\beta}_2 d^*\right)\right)\left(d-d^*\right) \\
\widehat{NIE}\left(d,d^*\right)^{lmi} &= \hat{\beta}_2\left(\hat{\gamma}_3 + \hat{\gamma}_4 d\right)\left(d-d^*\right) \\
\widehat{ATE}\left(d,d^*\right)^{lmi} &= \left(\left(\hat{\gamma}_2 + \hat{\gamma}_4\left(\hat{\beta}_0 + \hat{\beta}_2 d^*\right)\right) + \hat{\beta}_2\left(\hat{\gamma}_3 + \hat{\gamma}_4 d\right)\right)\left(d-d^*\right) \\
\widehat{CDE}\left(d,d^*,m\right)^{lmi} &= \left(\hat{\gamma}_2 + \hat{\gamma}_4 m\right)\left(d-d^*\right),
\end{aligned} \tag{3.34}$$

where the "lmi" superscript indicates that these estimators are based on linear models with an exposure-mediator interaction and possibly interactions of the exposure and mediator with the mean-centered confounders as well.

Using data from the NLSY, Table 3.2 reports point estimates for the total, direct, and indirect effects of college attendance on depression, as mediated by unemployment status. These estimates were computed using linear models for both the mediator and outcome, and they adjust for a vector of baseline confounders, including measures of race, gender, parental education, parental occupation, family size, family income, and scores on the Armed Forces Qualification Test (AFQT), which respondents were administered when they were in high school.

The estimates in the first column of Table 3.2 come from additive linear models equivalent to those in Equations 3.27 and 3.28. The estimates in the second column come from an additive linear model for the mediator and then a linear model for the outcome that includes an exposure-mediator interaction term, as in Equations 3.30 and 3.31. The estimates in the third column additionally include all two-way interactions of the confounders with the exposure and mediator, similar to the specifications in Equations 3.32 and 3.33. When computing the estimates reported in the second and third columns, the baseline confounders were centered around their sample means. All point estimates were constructed by fitting these models to the NLSY using OLS and then by substituting their coefficient estimates into Equations 3.29 and 3.34, as appropriate. Links to the code and data used for this analysis are provided in the table footnote.

Overall, these estimates do not suggest a very important role for unemployment in mediating the effect of education on depression at midlife. For example, consider the effect estimates from our most flexible model specifications, which are reported in the third column of the table. The estimate for the $ATE(1, 0)$ suggests that if sample members had, versus had not, attended college by age 22, their CES-D scores at age 40 would have been 0.129 standard deviations lower, on average. Our estimates for the natural and controlled direct effects are broadly similar to the total effect. Specifically, the estimate for the $NDE(1, 0)$ suggests that attending college would have reduced CES-D scores at age 40 by 0.119 standard deviations, on average, even if individuals had experienced the level of unemployment between age 35-39 that would have occurred for them had they not completed college. Similarly, the estimate for the $CDE(1, 0, 0)$ suggests that CES-D scores would have declined by 0.111 standard deviations as a result of attending college, even if everyone had avoided unemployment. By contrast, the estimate for the $NIE(1, 0)$ is close to zero, which suggests that if sample members were to experience the level of unemployment that would have occurred for them had they attended college, rather than the level of unemployment that would have occurred had they not attended college, their level of depression at midlife would barely change at all. This pattern of results is generally consistent across model specifications, although the effect estimates from our most flexible specification including exposure-mediator, confounder-exposure, and confounder-mediator interactions are more pronounced.

To conclude, total, direct, and indirect effects can be estimated using linear models for the mediator and outcome fit to sample data by the method of least squares. These estimators are consistent for their target estimands provided that the assumptions required for identification are satisfied and provided that the models used for estimation are correctly specified. In many analyses of mediation, researchers adopt specifications for these models that are additive in the predictors, as with Equations 3.27 and 3.28, but this is quite restrictive, raising the prospect of misspecification bias. Fortunately, it is not difficult to estimate the effects of interest using less restrictive models that allow for exposure-mediator interaction and effect moderation across levels of the baseline confounders. Thus, despite its simplicity, estimation with linear models is more flexible than it may at first appear, and this approach should remain in the toolkit of any analyst interested in causal mediation.

Nevertheless, models that are linear in the parameters may not perform very well in certain applications, no matter how flexibly they are specified. In particular, when the mediator or outcome is binary, ordinal, or a count, any linear model is likely incorrect and may provide poor estimates for the true but unknown conditional expected values, leading to misspecification bias. For example, in our analysis of the NLSY, the mediator of interest is binary, and thus the estimates in Table 3.2 are all based on linear probability models, whose limitations are well known (Agresti 2012; Aldrich and Nelson 1984). Although there are definitely situations where a linear model can provide a reasonable approximation for the conditional expected value of a binary, ordinal, or count variable, they are best suited for applications in which both the mediator and outcome are unbounded and possess equal-interval scaling. In the next section, we introduce a very general approach to estimation for total, direct, and indirect effects that can be implemented using a broad class of both linear and nonlinear models, which may perform better when analyzing causal mediation with discrete mediators and outcomes.

3.5.2 Estimation via Simulation

In this section, we explain how to estimate total, direct, and indirect effects using a simulation estimator that can be employed with generalized linear models (GLMs; Imai et al. 2010a,b, 2011). A GLM consists of three components: a distribution model for the dependent variable that falls in the exponential family of distributions, a linear predictor (i.e., a function of the predictors that is linear in the parameters), and an invertible link function that connects the dependent variable to the linear predictor. They are fit by the method of maximum likelihood, which selects estimates for model parameters by finding values that maximize the probability of observing the sample data given the model (Fox 2015; McCullagh 1989).

The class of GLMs is broad and subsumes many different models. It includes conventional linear regression as a special case, and it also includes a number of nonlinear models that are commonly employed in social science research, such as logit, probit, and Poisson regression, among others. We use the terms "linear" and "nonlinear" to distinguish between models that are either linear or nonlinear in the parameters, specifically. Models that are linear in the parameters may still include nonlinear functions of the predictors, as when a logarithmic or power transformation is applied to a covariate before including it in a conventional linear regression. They may not, however, include nonlinear functions of the model parameters.

The *simulation estimator* is implemented in series of steps. First, GLMs for the mediator and outcome are fit to the sample data. Next, the mediator model is used to simulate values for the mediator under different exposures. Then, the outcome model is used to simulate values for the outcome under different exposures and under different simulated values for the mediator. Finally, the simulated outcomes are used to construct estimates for the effects of interest. Specifically, when estimating the $ATE\left(d,d^*\right)$, $NDE\left(d,d^*\right)$, and $NIE\left(d,d^*\right)$, this procedure is implemented as follows:

1. **Fit models for the mediator and outcome.** That is, fit a GLM for the mediator given the baseline confounders and the exposure, denoted by $g\left(M|C,D\right)$, and then fit another GLM for the outcome given the baseline confounders, the exposure, and the mediator, denoted by $h\left(Y|C,D,M\right)$. Let $\hat{g}\left(M|C,D\right)$ and $\hat{h}\left(Y|C,D,M\right)$ denote these models with their parameters estimated by maximum likelihood.

2. **Simulate potential values for the mediator.** For every individual in the sample, simulate J copies of $M\left(d^*\right)$ from $\hat{g}\left(M|C,d^*\right)$, and then simulate another J copies of $M\left(d\right)$ from $\hat{g}\left(M|C,d\right)$. Let $\tilde{M}_j\left(d^*\right)$ and $\tilde{M}_j\left(d\right)$ denote the simulated values of the mediator for each simulation $j = 1, 2, \ldots, J$.

3. **Simulate potential outcomes.** For every individual in the sample and for each simulated value of the mediator, simulate one copy of $Y\left(d,M\left(d\right)\right)$ from $\hat{h}\left(Y|C,d,\tilde{M}_j\left(d\right)\right)$, one copy of $Y\left(d,M\left(d^*\right)\right)$ from $\hat{h}\left(Y|C,d,\tilde{M}_j\left(d^*\right)\right)$, and one copy of $Y\left(d^*,M\left(d^*\right)\right)$ from $\hat{h}\left(Y|C,d^*,\tilde{M}_j\left(d^*\right)\right)$. Let $\tilde{Y}_j\left(d,M\left(d\right)\right)$, $\tilde{Y}_j\left(d,M\left(d^*\right)\right)$, and $\tilde{Y}_j\left(d^*,M\left(d^*\right)\right)$ denote these simulated outcomes for each simulation $j = 1, 2, \ldots, J$.

4. **Compute effect estimates.** Estimators for the total, natural direct, and natural indirect effects are given by the following functions of the simulated outcomes:

$$\begin{aligned}
\widehat{NDE}\left(d,d^*\right)^{sim} &= \frac{1}{nJ}\sum\sum_{j}\left(\tilde{Y}_j\left(d,M\left(d^*\right)\right) - \tilde{Y}_j\left(d^*,M\left(d^*\right)\right)\right) \\
\widehat{NIE}\left(d,d^*\right)^{sim} &= \frac{1}{nJ}\sum\sum_{j}\left(\tilde{Y}_j\left(d,M\left(d\right)\right) - \tilde{Y}_j\left(d,M\left(d^*\right)\right)\right) \\
\widehat{ATE}\left(d,d^*\right)^{sim} &= \frac{1}{nJ}\sum\sum_{j}\left(\tilde{Y}_j\left(d,M\left(d\right)\right) - \tilde{Y}_j\left(d^*,M\left(d^*\right)\right)\right), \qquad (3.35)
\end{aligned}$$

where the inner sum is taken over the J simulations and the outer sum is taken over the n sample members. The "sim" superscript in these expressions denotes that they are simulation estimators.

In step 1 of this estimation procedure, we first fit a GLM for the mediator with the baseline confounders and the exposure as predictors, and then we fit another GLM for the outcome with the confounders, exposure, and mediator as predictors. For example, if the mediator and outcome are both binary, these models might be logit or probit regressions. If the mediator is continuous but the outcome binary, the GLM for the mediator might be a normal linear model, while the GLM for the outcome might be a logit or probit regression. In general, any combination of different GLMs for the mediator and outcome can be used to implement the simulation estimator. Whichever models are selected, what matters is that they are correctly specified.

In step 2, we use the fitted model for the mediator to simulate values under different exposures. Specifically, for each sample member, we first replace their exposure with the value d^*, while leaving their observed values on the baseline confounders intact, and then we select J Monte Carlo samples of the mediator from our fitted model with the predictors set at these levels. The resulting simulated values are denoted by $\tilde{M}_j\left(d^*\right)$ for $j = 1, 2, \ldots, J$. They represent estimates of the potential value of the mediator under exposure d^*. Next, we replace each sample member's exposure with the value d, while still leaving their baseline confounders undisturbed, and then we select another set of J Monte Carlo samples from our fitted model with the predictors now set at these levels. The resulting simulated values are denoted by $\tilde{M}_j\left(d\right)$ for $j = 1, 2, \ldots, J$, and they represent estimates of the potential value of the mediator under exposure d.

In step 3, we use the fitted model for the outcome to simulate values under different exposures and under different simulated values for the mediator. Specifically, for each sample member, we begin by replacing their exposure with the value d and their mediator with one of its simulated values $\tilde{M}_j\left(d\right)$, while leaving their observed values on the baseline confounders intact. With the predictors set at these levels, we then select one Monte Carlo sample of the outcome from our fitted model. Repeating this procedure for each simulated value of the mediator yields a corresponding set of simulated values for the outcome, which are denoted by $\tilde{Y}_j\left(d, M\left(d\right)\right)$ for $j = 1, 2, \ldots, J$. These simulated values represent estimates of a sample member's potential outcome under exposure d and under the level of the mediator that they would have experienced under exposure to d.

Next, with each sample member's exposure still set at d and their confounders still undisturbed, we replace their mediator with one of its simulated values under the alternative level of the exposure–that is, we reset their mediator at a value given by $\tilde{M}_j\left(d^*\right)$. With the predictors now set at these levels, we then select one Monte Carlo sample of the outcome from our fitted model. Repeating this procedure for each value of $\tilde{M}_j\left(d^*\right)$ yields a corresponding set of simulated values for the outcome, denoted by $\tilde{Y}_j\left(d, M\left(d^*\right)\right)$ for $j = 1, 2, \ldots, J$. These simulated values represent estimates of a sample member's cross-world potential outcome.

Finally, we set each sample member's exposure at d^* and their mediator at a simulated value given by $\tilde{M}_j\left(d^*\right)$, while still leaving their observed values on the baseline confounders intact. We then select another Monte Carlo sample from our fitted model for the outcome, repeating this procedure for each simulated value of the mediator. The resulting set of simulated outcomes are denoted by $\tilde{Y}_j\left(d^*, M\left(d^*\right)\right)$ for $j = 1, 2, \ldots, J$, and they represent estimates of a sample member's potential outcome under exposure d^* and under the level of the mediator that they would have experienced under exposure to d^*. At the conclusion of step 3, we have thus obtained J simulated values for each of the potential outcomes that define the total, direct, and indirect effects of interest.

In step 4 of the estimation procedure, we simply take differences between the simulated outcomes, and then we average them over simulations and over sample members to produce estimates for the effects of interest. For example, to compute an estimate of $NDE\left(d, d^*\right)$, we would take the difference between $\tilde{Y}_j\left(d, M\left(d^*\right)\right)$ and $\tilde{Y}_j\left(d^*, M\left(d^*\right)\right)$ for each simulation and for each sample member, and then we would average all of these differences together. Estimates for the $NIE\left(d, d^*\right)$ and $ATE\left(d, d^*\right)$ are computed in essentially the same way but with different contrasts between the simulated outcomes.

To better understand the logic of the simulation approach, consider its connection to the identification formulas described earlier. For example, under assumptions (c.i) to (c.vi), the identification formula for the marginal expected value of the cross-world potential outcome $Y\left(d, M\left(d^*\right)\right)$ can be expressed as follows:

$$\begin{aligned}\mathbb{E}\left[Y\left(d, M\left(d^*\right)\right)\right] &= \sum_{m,c} \mathbb{E}\left[Y|c, d, m\right] P\left(m|c, d^*\right) P\left(c\right) \\ &= \sum_{c} \Big(\sum_{y,m} yP\left(y|c, d, m\right) P\left(m|c, d^*\right)\Big) P\left(c\right). \qquad (3.36)\end{aligned}$$

For continuous data, the probability-weighted sums are replaced by density-weighted integrals. In either case, the identification formula depends on the conditional distributions of the mediator and outcome, denoted by $P\left(m|c, d^*\right)$ and $P\left(y|c, d, m\right)$ in the expression above. The simulation approach models these distributions parametrically and then generates Monte Carlo samples from them. By averaging these samples across simulations, we approximate the sum over y and m in Equation 3.36, while averaging the resulting quantities again across sample members approximates the sum over c, now using the empirical distribution of the confounders to estimate $P\left(c\right)$ rather than a parametric model. Taking the mean of Monte Carlo samples provides a simple way to approximate potentially complex sums or integrals in an identification formula, particularly when evaluating them directly is challenging due to the use of nonlinear or otherwise complex models for the mediator and outcome.

To illustrate this estimation procedure more concretely, consider the following example. Suppose that we have sample data on a continuous exposure D, a binary mediator M, and an outcome Y equal to a discrete count of events. How might we implement the simulation estimator to decompose the average total effect of exposure on the outcome into natural direct and indirect components in this scenario?

We would first need to specify and fit GLMs for the mediator and outcome. Because M is binary and Y a discrete count, we might consider a logit model and a Poisson model, respectively, for the mediator and outcome. These models can be formally expressed as follows:

$$g\left(M|c, d\right) = P\left(M = m|c, d\right) = Bern\left(p = logit^{-1}\left(\beta_0 + \beta_1^T c + \beta_2 d\right)\right) \qquad (3.37)$$

$$h\left(Y|c, d, m\right) = P\left(Y = y|c, d, m\right) = Pois\left(\lambda = exp\left(\gamma_0 + \gamma_1^T c + \gamma_2 d + m\left(\gamma_3 + \gamma_4 d\right)\right)\right), \qquad (3.38)$$

where $exp\left(\cdot\right)$ is the exponential function and $logit^{-1}\left(\cdot\right) = \frac{exp(\cdot)}{1+exp(\cdot)}$ is the inverse of the logit, or log odds, function.

In Equation 3.37, $Bern\left(p\right)$ denotes the Bernoulli probability distribution, which is a discrete probability distribution for a binary random variable that takes the value 1 with probability p and the value 0 with probability $1-p$. This equation is a GLM because it includes a distribution model – the Bernoulli distribution – in the exponential family, because it has a linear predictor given by $\beta_0 + \beta_1^T c + \beta_2 d$, and because it has an invertible link function, given by $logit^{-1}\left(\cdot\right)$, that connects the linear predictor with the probability of the mediator.

Similarly, in Equation 3.38, $Pois\left(\lambda\right)$ denotes the Poisson probability distribution. The Poisson distribution is another discrete probability distribution for the number of events in a fixed interval of time or space, where the probability mass over events is governed by the parameter λ. Equation 3.38 is also a GLM because it includes a distribution model – the Poisson distribution – in the exponential family, because it has a linear predictor given by $\gamma_0 + \gamma_1^T c + \gamma_2 d + m\left(\gamma_3 + \gamma_4 d\right)$, and because it connects this linear predictor to the probability of the outcome by way of an invertible link function given by $exp\left(\cdot\right)$.

After fitting these models by the method of maximum likelihood, we would next simulate values for the mediator under different levels of the exposure. To this end, we would replace each sample member's exposure with the value d^* and select J Monte Carlo samples of $\tilde{M}_j\left(d^*\right)$ from a Bernoulli distribution with $\hat{p} = logit^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1^T C + \hat{\beta}_2 d^*\right)$, where "hats" here denote maximum likelihood estimates. We would also replace each sample member's exposure with the value d and select another J Monte Carlo samples of $\tilde{M}_j\left(d\right)$ from a Bernoulli distribution with $\hat{p} = logit^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1^T C + \hat{\beta}_2 d\right)$.

With simulated values of the mediator in hand, we would then turn to simulating values for the outcome. Specifically, for each sample member and for each simulated value j of the mediator, we would now select one Monte Carlo sample of $\tilde{Y}_j\left(d, M\left(d\right)\right)$ from a Poisson distribution with $\hat{\lambda} = exp\left(\hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_2 d + \tilde{M}_j\left(d\right)\left(\hat{\gamma}_3 + \hat{\gamma}_4 d\right)\right)$, one Monte Carlo sample of $\tilde{Y}_j\left(d, M\left(d^*\right)\right)$ from a Poisson distribution with $\hat{\lambda} = exp\left(\hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_2 d + \tilde{M}_j\left(d^*\right)\left(\hat{\gamma}_3 + \hat{\gamma}_4 d\right)\right)$, and one Monte Carlo sample of $\tilde{Y}_j\left(d^*, M\left(d^*\right)\right)$ from a Poisson distribution with $\hat{\lambda} = exp\left(\hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_2 d^* + \tilde{M}_j\left(d^*\right)\left(\hat{\gamma}_3 + \hat{\gamma}_4 d^*\right)\right)$.

Finally, to conclude the procedure, we would plug the simulated outcomes into our expressions for $\widehat{NDE}\left(d, d^*\right)^{sim}$, $\widehat{NIE}\left(d, d^*\right)^{sim}$, and $\widehat{ATE}\left(d, d^*\right)^{sim}$. Solving these expressions would yield consistent estimates for the effects of interest provided that our identification assumptions are satisfied and that our distribution models in Equations 3.37 and 3.38 are correct.

A similar procedure can be used to estimate controlled direct effects, but when targeting the $CDE\left(d, d^*, m\right)$, we need only predict the outcome from our fitted model rather than simulate it using many different Monte Carlo samples. This simplification is possible because, with only baseline confounding, the identification formula for the controlled direct effect does not require averaging over the distribution of a mediator, as with natural effects; rather, it just involves averaging some conditional means of the outcome over the marginal distribution of the confounders. This can be achieved using the empirical distribution of the confounders in the sample data, without any Monte Carlo draws from a separate parametric model.

Specifically, an *imputation estimator* for the controlled direct effect can be implemented as follows:

1. **Fit a model for the outcome.** That is, fit a GLM for the outcome given the baseline confounders, the exposure, and the mediator, denoted by $h\left(Y|C, D, M\right)$. Let $\hat{h}\left(Y|C, D, M\right)$ represent the fitted model with parameters estimated by maximum likelihood.

2. **Impute potential outcomes.** For every individual in the sample, obtain their predicted value for $Y\left(d^*, m\right)$ using $\hat{h}\left(Y|C, d^*, m\right)$, and then obtain their predicted value for $Y\left(d, m\right)$ using $\hat{h}\left(Y|C, d, m\right)$. Let $\hat{Y}\left(d^*, m\right)$ and $\hat{Y}\left(d, m\right)$ denote these predicted values for each sample member.

3. **Compute an effect estimate.** An estimator for the controlled direct effect is given by the following function of the predicted outcomes:

$$\widehat{CDE}\left(d, d^*, m\right)^{imp} = \frac{1}{n}\sum\left(\hat{Y}\left(d, m\right) - \hat{Y}\left(d^*, m\right)\right), \tag{3.39}$$

   where the sum is taken over the n sample members and the "imp" superscript denotes that this expression is an imputation estimator.

In step 1 of this estimation procedure, we first fit a GLM for the outcome with the confounders, exposure, and mediator as predictors. As with the simulation estimator for total, natural direct, and natural indirect effects, the imputation estimator for controlled direct effects can be implemented with any GLM for the outcome. If our outcome were a discrete count, as in our hypothetical example above, we might consider a Poisson model like Equation 3.38 for this variable.

In step 2, we use our fitted model for the outcome to obtain predicted values under different exposures. For each sample member, we first replace their exposure and mediator with the values d^* and m, respectively, while leaving their observed values on the baseline confounders intact. We then obtain predicted values, denoted by $\hat{Y}\left(d^*, m\right)$, from our fitted model with the predictors set at these levels. Next, we replace each sample member's exposure with the value d, but we still leave their baseline confounders undisturbed and their mediator set at level m. We then obtain another set of predicted values, denoted by $\hat{Y}\left(d, m\right)$, from our fitted model with the predictors now set at these levels. With a Poisson model as in Equation 3.38, for example, the predicted values would be given by $\hat{Y}\left(d^*, m\right) = exp\left(\hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_2 d^* + m\left(\hat{\gamma}_3 + \hat{\gamma}_4 d^*\right)\right)$ and $\hat{Y}\left(d, m\right) = exp\left(\hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_2 d + m\left(\hat{\gamma}_3 + \hat{\gamma}_4 d\right)\right)$.

In step 3, we compute differences between these predicted values and then average them over sample members to estimate the controlled direct effect. That is, we plug the predicted outcomes into the expression for $\widehat{CDE}\left(d, d^*, m\right)^{imp}$ and solve. This procedure yields consistent estimates for controlled direct effects provided that our identification assumptions are satisfied and our GLM for the outcome is correctly specified.

Using data from the NLSY, Table 3.3 reports another set of point estimates for the effects of college attendance on depression, as mediated by unemployment status. For the total, natural direct, and natural indirect effects, these results come from the simulation estimator, which we implemented using a logit model for the mediator and a normal linear model for the outcome. Specifically, the estimates in the first column of the table are based on a logit model for $g\left(M|c, d\right)$, where the mediator is assumed to follow a Bernoulli distribution with a conditional probability given by $p = logit^{-1}\left(\beta_0 + \beta_1^T c + \beta_2 d\right)$. They are also based on a normal linear model for $h\left(Y|c, d, m\right)$, where the outcome is assumed to be normally distributed with a conditional mean given by $\mu = \gamma_0 + \gamma_1^T c + \gamma_2 d + \gamma_3 m$ and a homoscedastic conditional variance given by the model's mean squared error. The results in the second column of the table are based on a similar set of models, but for these estimates, we additionally included all two-way interactions between the baseline confounders and exposure in the logit model for the mediator as well as all two-way interactions among the baseline confounders, exposure, and mediator in the normal linear model for the outcome. Both sets of estimates were constructed by fitting these models to the NLSY and then by using the fitted models to generate $J = 2000$ simulations. For the controlled direct effect, the results in Table 3.3 come from the imputation estimator, which we implemented using the same set of outcome models as outlined previously for the simulation approach. Links to the code and data used to produce these results are provided in the table footnote.

Estimates computed using the simulation approach and a logit rather than linear model for the mediator are very similar to those reported previously in Table 3.2, which were based exclusively on linear models for both the mediator and outcome. As with our previous estimates based on linear models, the simulation estimates also do not suggest a very important role for unemployment in mediating the effect of education on depression. Specifically, simulation estimates for the total and direct effects suggest that attending college would reduce depression by a nontrivial margin. By contrast, simulation estimates for the natural indirect effect are close to zero, indicating a minimal explanatory role for unemployment in the causal process connecting education with later life depression. This pattern of results is consistent across the different

Table 3.3: Total, Direct, and Indirect Effects of College Attendance on CES-D Scores as Estimated from the NLSY using the Simulation and Imputation Approach.

Estimand	Point Estimates	
	Additive Logit Model for M (LogitMod); Normal, Additive Linear Model for Y (LinMod)	LogitMod with $C \times D$ Interaction Terms; Normal LinMod with $D \times M$, $C \times D$ and $C \times M$ Interaction Terms
$ATE(1,0)$	$-.081$	$-.132$
$NDE(1,0)$	$-.072$	$-.120$
$NIE(1,0)$	$-.009$	$-.012$
$CDE(1,0,0)$	$-.072$	$-.111$

Note: Estimates are expressed in standard deviation units. D denotes college attendance, M unemployment status, C the baseline confounders, and Y CES-D scores. Results are based on $J = 2000$ simulations. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/table_3-3`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

model specifications we used to implement the simulation estimator, although effect estimates from the more flexible of these specifications are more pronounced.

To summarize, total, direct, and indirect effects can be estimated by fitting GLMs to sample data and then by using these models to simulate or impute the potential outcomes that define the effects of interest. If the assumptions required to identify these effects are satisfied, and if the models assumed for the mediator and outcome are correctly specified, then the simulation approach yields estimators that converge to their target estimands as both the sample size n and number of Monte Carlo simulations J increase indefinitely. If, however, the identification assumptions are not satisfied, then this approach to estimation will suffer from systematic bias. Moreover, even when these identification assumptions are met, the simulation approach may still yield biased and inconsistent estimates if either of the GLMs for the mediator or the outcome is incorrectly specified. For example, if the GLMs adopt the wrong distribution model (e.g., Poisson versus negative binomial), or the wrong link function (e.g., logistic versus probit), or the wrong linear predictor (e.g., one with versus without interaction terms among the exposure, mediator, and confounders), then the estimates may suffer from systematic errors, even when the requisite confounding, consistency, and positivity assumptions are all satisfied and the effects of interest can therefore be identified from the observed data nonparametrically.

Unlike the other estimators we considered in prior sections, the simulation estimator suffers from not one but two sources of random variability: sampling error as a result of observing only a random subset of individuals from the target population, and simulation error as a result of taking a finite number of Monte Carlo draws from the fitted distribution model. The magnitude of random variability due to simulation error depends on the number of Monte Carlo samples J. The larger is J, the smaller is the degree of simulation error in the resulting estimates. In practice, $10^3 \leq J \leq 10^4$ should yield estimates whose random variability is not meaningfully inflated by simulation error, above and beyond error due to random sampling from the target population. The only cost associated with using a larger J is the longer computation time and greater memory capacity needed to produce, store, and analyze the additional simulations. With most modern computers and empirical applications, these costs will be trivial for the recommend number of Monte Carlo samples.

The simulation and imputation estimators outlined in this section are extremely flexible because GLMs include a broad class of models that can accommodate many different types of mediators and outcomes, many different types of probability distributions for these variables, and many different types of relationships among the predictors, mediator, and outcome, depending on the link function and specification of the linear predictor. In fact, although we have focused on implementing the simulation and imputation estimators with GLMs, these methods can actually be implemented with virtually any parametric or semi-parametric distribution model for the mediator and outcome (Imai et al. 2010a). They therefore represent a highly general and adaptable approach to model-based estimation of total, direct, and indirect effects. Compared with the estimators outlined in Section 3.5.1, which relied exclusively on linear regression models, the generality and adaptability of the simulation estimator make it better suited for applications with binary, ordinal or count variables.

3.5.3 Estimation with Inverse Probability Weights

In this section, we explain how to estimate total, direct, and indirect effects using inverse probability weights (Hong 2015; Nguyen et al. 2023; Tchetgen Tchetgen 2013; VanderWeele 2009b). In contrast to the estimators outlined previously, which are based on models for the mediator and outcome, weighting estimators are implemented mainly with models for the exposure, although in some cases they require a model for the mediator as well. These models are used to construct a set of weights that transform the empirical distribution of the sample data in ways that emulate different hypothetical experiments, and then the effects of interest are estimated by comparing the mean of the outcome across differently weighted samples.

Specifically, a set of *weighting estimators* for the $ATE\left(d, d^*\right)$, $NDE\left(d, d^*\right)$, and $NIE\left(d, d^*\right)$ can be constructed through the following series of steps. First, two separate GLMs for the exposure are fit to the sample data. One of these GLMs includes only the baseline confounders as predictors, while the other includes both the confounders and the mediator. Next, the fitted models are used to estimate exposure probabilities under each set of predictors, and then weights are constructed based on these probabilities to balance the confounders and mediator across levels of the exposure in different ways. Finally, the effects of interest are estimated by comparing outcome means across the weighted samples.

When estimating total, natural direct, and natural indirect effects, this procedure is implemented as follows:

1. **Fit models for the exposure.** That is, fit a GLM for the exposure given the baseline confounders, denoted by $f\left(D|C\right)$. Then, fit another GLM for the exposure given both the baseline confounders and the mediator, denoted by $s\left(D|C,M\right)$. Let $\hat{f}\left(D|C\right)$ and $\hat{s}\left(D|C,M\right)$ denote these models with their parameters estimated by maximum likelihood.

2. **Compute predicted probabilities of exposure.** For each sample member, use $\hat{f}\left(D|C\right)$ to predict the probability of exposure to d conditional on their observed values for the baseline confounders. Next, use $\hat{f}\left(D|C\right)$ to predict the probability of exposure to d^* conditional on the baseline confounders. Then, use $\hat{s}\left(D|C,M\right)$ to predict the probability of exposure to d given a sample member's observed values on both the baseline confounders and the mediator. Finally, use $\hat{s}\left(D|C,M\right)$ to predict the probability of exposure to d^* conditional on the baseline confounders and the mediator. Let $\hat{P}\left(d|C\right)$, $\hat{P}\left(d^*|C\right)$, $\hat{P}\left(d|C,M\right)$, and $\hat{P}\left(d^*|C,M\right)$ denote each of these predicted probabilities in turn.

3. **Construct inverse probability weights.** Among sample members for whom $D = d^*$, compute a set of inverse probability weights given by $\hat{w}_1 = {}^1/_{\hat{P}\left(d^*|C\right)}$. Then, among sample members for whom

$D = d$, compute two additional sets of inverse probability weights given by $\hat{w}_2 = {}^{1}/_{\hat{P}(d|C)}$ and $\hat{w}_3 = {}^{\hat{P}(d^*|C,M)}/_{\hat{P}(d|C,M)\hat{P}(d^*|C)}$.

4. **Compute effect estimates.** Estimators for the total, natural direct, and natural indirect effects are given by the following contrasts between weighted means of the observed outcome:

$$\begin{aligned}
\widehat{NDE}\left(d, d^*\right)^{ipw} &= \frac{\sum I\left(D=d\right)\hat{w}_3 Y}{\sum I\left(D=d\right)\hat{w}_3} - \frac{\sum I\left(D=d^*\right)\hat{w}_1 Y}{\sum I\left(D=d^*\right)\hat{w}_1} \\
\widehat{NIE}\left(d, d^*\right)^{ipw} &= \frac{\sum I\left(D=d\right)\hat{w}_2 Y}{\sum I\left(D=d\right)\hat{w}_2} - \frac{\sum I\left(D=d\right)\hat{w}_3 Y}{\sum I\left(D=d\right)\hat{w}_3} \\
\widehat{ATE}\left(d, d^*\right)^{ipw} &= \frac{\sum I\left(D=d\right)\hat{w}_2 Y}{\sum I\left(D=d\right)\hat{w}_2} - \frac{\sum I\left(D=d^*\right)\hat{w}_1 Y}{\sum I\left(D=d^*\right)\hat{w}_1},
\end{aligned} \tag{3.40}$$

where the "ipw" superscript denotes that they are based on inverse probability weights. In these expressions, $I\left(\cdot\right)$ is an indicator function equal to 1 when its argument is true, and 0 otherwise. Thus, ${}^{\sum I(D=d^*)\hat{w}_1 Y}/_{\sum I(D=d^*)\hat{w}_1}$ is a weighted mean of the outcome Y among sample members for whom $D = d^*$, with weights given by $\hat{w}_1 = {}^{1}/_{\hat{P}(d^*|C)}$. Similarly, ${}^{\sum I(D=d)\hat{w}_2 Y}/_{\sum I(D=d)\hat{w}_2}$ and ${}^{\sum I(D=d)\hat{w}_3 Y}/_{\sum I(D=d)\hat{w}_3}$ are weighted means of the outcome among sample members for whom $D = d$, with weights given by $\hat{w}_2 = {}^{1}/_{\hat{P}(d|C)}$ and $\hat{w}_3 = {}^{\hat{P}(d^*|C,M)}/_{\hat{P}(d|C,M)\hat{P}(d^*|C)}$, respectively.

In step 1 of this estimation procedure, we first fit a GLM for the exposure with only the baseline confounders as predictors. Then, we fit another GLM for the exposure but now with both the confounders and the mediator as predictors. We use $\hat{f}\left(D|C\right)$ and $\hat{s}\left(D|C, M\right)$ to denote the maximum likelihood fit for each of these models in turn. If the exposure is binary, for example, both models might be logit or probit regressions.

In step 2, we use these fitted models to predict the probability that a sample member would experience each alternative level of the exposure, d versus d^*, that together define the effects of interest. Specifically, we first use $\hat{f}\left(D|C\right)$ to predict the probability that each sample member is exposed to d given their observed values on the baseline confounders. We also use this model to predict the probability that each sample member is exposed to d^* conditional on the confounders. These predicted probabilities are denoted by $\hat{P}\left(d|C\right)$ and $\hat{P}\left(d^*|C\right)$, respectively. Next, we use $\hat{s}\left(D|C, M\right)$ to predict the probability that each sample member is exposed to d given their observed values on both the confounders and the mediator. We also use this model to predict the probability that each sample member is exposed to d^* conditional on their observed values for the confounders and mediator. These predicted probabilities are denoted by $\hat{P}\left(d|C, M\right)$ and $\hat{P}\left(d^*|C, M\right)$, respectively.

In step 3, we construct several sets of inverse probability weights. The first set of weights is given by $\hat{w}_1 = {}^{1}/_{\hat{P}(d^*|C)}$, which represents the inverse probability of exposure to d^* given the baseline confounders. Among sample members for whom $D = d^*$, weighting by $\hat{w}_1$ transforms the distribution of the confounders in this subsample to resemble their distribution in the total sample taken as a whole. In other words, weighting by $\hat{w}_1$ creates a pseudosample in which assignment to exposure d^* appears to have occurred at random with respect to the baseline confounders. Weighting by $\hat{w}_1$ achieves this transformation by giving greater weight to sample members with confounder values that make them less likely to experience the exposure d^*, and conversely, by giving lesser weight to sample members with confounder values that make them more likely to experience this exposure. By up-weighting those with confounder values that are underrepresented, and by down-weighting those with confounder values that are over-represented, the inverse probability weights bring the distribution of the confounders among sample members for whom $D = d^*$ back into balance with the total sample.

Similarly, the second set of weights is given by $\hat{w}_2 = {}^1/\hat{P}(d|C)$, which represents the inverse probability of exposure to d given the baseline confounders. Among sample members for whom $D = d$, weighting by $\hat{w}_2$ transforms the distribution of the confounders in this subsample to resemble their distribution in the total sample considered altogether. In other words, weighting by $\hat{w}_2$ creates a pseudosample in which assignment to exposure d appears to have occurred at random with respect to the baseline confounders. Weighting by $\hat{w}_2$ achieves this transformation by giving greater weight to sample members with confounder values that are underrepresented, and lesser weight to sample members with confounder values that are over-represented, among those for whom $D = d$. In this way, inverse probability weighting shifts the distribution of the confounders in the subsample exposed to d so that it mirrors the distribution observed in the total sample as a whole.

The third set of weights is given by $\hat{w}_3 = \hat{P}(d^*|C,M)/\hat{P}(d|C,M)\hat{P}(d^*|C)$, which is the product of an odds ratio $\hat{P}(d^*|C,M)/\hat{P}(d|C,M)$ and an inverse probability ${}^1/\hat{P}(d^*|C)$. The odds ratio is equal to the probability of exposure to d^* given a sample member's observed values on both the confounders and the mediator divided by the probability of exposure to d conditional on these same variables. Among sample members for whom $D = d$, weighting by the odds ratio $\hat{P}(d^*|C,M)/\hat{P}(d|C,M)$ transforms the distribution of both the confounders and the mediator in this subsample to resemble their distribution among the other subsample of individuals for whom $D = d^*$. Additionally weighting by the inverse probability ${}^1/\hat{P}(d^*|C)$ further transforms the distribution of the confounders among those with $D = d$ so that it resembles the confounder distribution in the total sample. Altogether, then, weighting by $\hat{w}_3$ creates a pseudosample in which assignment to exposure d appears to have occurred at random with respect to the baseline confounders and in which the distribution of the mediator mirrors that found among sample members for whom $D = d^*$. In other words, weighting by $\hat{w}_3$ creates a *cross-world* pseudosample where individuals exposed to d appear to have the distribution of M observed among those exposed to d^* and the distribution of C observed among the total sample as a whole.

In step 4 of this estimation procedure, we compute the mean of the observed outcome in each of these weighted pseudosamples, and then we take differences between them to estimate total, natural direct, and natural indirect effects. Specifically, because weighting the subsample for whom $D = d^*$ by $\hat{w}_1$ creates a pseudosample in which exposure to d^* appears to have occurred at random but the distribution of M is unchanged, the mean of the outcome in this pseudosample, which is given by $\sum I(D=d^*)\hat{w}_1 Y/\sum I(D=d^*)\hat{w}_1$, yields an estimate for $\mathbb{E}\left[Y\left(d^*, M\left(d^*\right)\right)\right]$. Similarly, because weighting the subsample for whom $D = d$ by $\hat{w}_2$ creates a pseudosample in which exposure to d appears to have occurred at random but the distribution of M is unchanged, the mean of the outcome in this pseudosample, which is given by $\sum I(D=d)\hat{w}_2 Y/\sum I(D=d)\hat{w}_2$, yields an estimate for $\mathbb{E}\left[Y\left(d, M\left(d\right)\right)\right]$. Finally, because weighting the subsample for whom $D = d$ by $\hat{w}_3$ creates a cross-world pseudosample with the distribution of M found among those with $D = d^*$ and the distribution of C observed among all sample members taken together, the mean of the outcome in this pseudosample, which is given by $\sum I(D=d)\hat{w}_3 Y/\sum I(D=d)\hat{w}_3$, yields an estimate for $\mathbb{E}\left[Y\left(d, M\left(d^*\right)\right)\right]$. Computing differences between these weighted means following Equation 3.40 gives consistent estimates for the effects of interest, provided that our identification assumptions are satisfied and that the models used to construct the weights are correctly specified.

In Appendix F, we outline an alternative approach to re-weighting the sample data when estimating natural direct and indirect effects. This alternative approach, known as ratio of mediator probability weighting (Hong et al. 2015; Hong 2015), is based on a model for the exposure and a model for the mediator rather than two separate models for the exposure with different sets of predictors. Ratio of mediator probability weighting is asymptotically equivalent to the weighting approach described previously if all the models used to construct the weights are correctly specified.

Similar procedures can be used to estimate controlled direct effects. When targeting the $CDE(d, d^*, m)$, GLMs for both the exposure and mediator are first fit to the sample data. Next, the fitted models are used to predict the probability that sample members experience their observed levels of the exposure and mediator. Then, based on these probabilities, a set of weights are constructed to transform the sample data so that the exposure and mediator appear to have been jointly randomized. Finally, the controlled direct effects of interest are estimated by comparing outcome means across differently weighted subsamples.

Specifically, inverse probability weighting for the $CDE(d, d^*, m)$ is implemented as follows:

1. **Fit models for the exposure and mediator.** That is, fit a GLM for the exposure given the baseline confounders, denoted by $f\,(D|C)$. Then, fit another GLM for the mediator given the confounders and the exposure, denoted by $g\,(M|C, D)$. Let $\hat{f}\,(D|C)$ and $\hat{g}\,(M|C, D)$ denote these models with their parameters estimated by maximum likelihood.

2. **Compute predicted probabilities.** That is, use $\hat{f}\,(D|C)$ to predict the probability that each sample member experiences their observed exposure given their baseline confounders. Let $\hat{P}\,(D|C)$ denote this set of predicted probabilities. Next, use $\hat{g}\,(M|C, D)$ to predict the probability that each sample member experiences their observed value of the mediator given their baseline confounders and exposure. Let $\hat{P}\,(M|C, D)$ denote these predictions.

3. **Construct inverse probability weights.** For all sample members, compute a set of inverse probability weights given by $\hat{w}_4 = {}^1\!/_{\hat{P}(M|C,D)\hat{P}(D|C)}$.

4. **Compute effect estimates.** An estimator for the controlled direct effect is given by the following contrast between weighted means of the observed outcome:

$$\widehat{CDE}\,(d, d^*, m)^{ipw} = \frac{\sum I\,(D = d, M = m)\,\hat{w}_4 Y}{\sum I\,(D = d, M = m)\,\hat{w}_4} - \frac{\sum I\,(D = d^*, M = m)\,\hat{w}_4 Y}{\sum I\,(D = d^*, M = m)\,\hat{w}_4}, \tag{3.41}$$

   where $I\,(\cdot)$ is an indicator function equal to 1 when its argument is true, and 0 otherwise, as above. In this expression, ${}^{\sum I(D=d,M=m)\hat{w}_4 Y}\!/_{\sum I(D=d,M=m)\hat{w}_4}$ is a weighted mean of the outcome Y among sample members for whom $D = d$ and $M = m$, with weights given by $\hat{w}_4 = {}^1\!/_{\hat{P}(m|C,d)\hat{P}(d|C)}$. Similarly, ${}^{\sum I(D=d^*,M=m)\hat{w}_4 Y}\!/_{\sum I(D=d^*,M=m)\hat{w}_4}$ is a weighted mean of the outcome among sample members for whom $D = d^*$ and $M = m$, with weights given by $\hat{w}_4 = {}^1\!/_{\hat{P}(m|C,d^*)\hat{P}(d^*|C)}$.

In step 1 of this estimation procedure, we first fit a GLM for the exposure with the baseline confounders as predictors, and then we fit another GLM for the mediator with the confounders and the exposure as predictors. If the exposure and mediator were both binary, for example, these models might be logit or probit regressions.

In step 2, we use these fitted models to predict the probability that sample members experience their observed values on the exposure and mediator. Specifically, we use our fitted model for the exposure, $\hat{f}\,(D|C)$, to predict the probability that each sample member experiences their observed level of the exposure given their baseline confounders. We then use our fitted model for the mediator, $\hat{g}\,(M|C, D)$, to predict the probability that each sample member experiences their observed level of the mediator given both their baseline confounders and their exposure. These predicted probabilities are denoted by $\hat{P}\,(D|C)$ and $\hat{P}\,(M|C, D)$, respectively.

In step 3, we construct a set of inverse probability weights. These weights are given by $\hat{w}_4 = {}^{1}/_{\hat{P}(M|C,D)\hat{P}(D|C)}$, which is the product of an inverse probability for the mediator ${}^{1}/_{\hat{P}(M|C,D)}$ and an inverse probability for the exposure ${}^{1}/_{\hat{P}(D|C)}$. Weighting the sample by $\hat{w}_4$ balances the distribution of the baseline confounders across levels of both the exposure and the mediator, or in other words, it creates a pseudosample in which the exposure and mediator appear to have been jointly randomized with respect to the confounders.

In step 4 of this estimation procedure, we compute an estimate for the controlled direct effect by comparing weighted means of the observed outcome across different subsamples. Specifically, because weighting the subsample for whom $D = d$ and $M = m$ by $\hat{w}_4$ creates a pseudosample in which exposure to both d and m appears to have occurred at random, the mean of the outcome in this pseudosample, which is given by $\sum {}^{I(D=d,M=m)\hat{w}_4 Y}/_{\sum I(D=d,M=m)\hat{w}_4}$, yields an estimate for $\mathbb{E}\left[Y\left(d, m\right)\right]$. Similarly, because weighting the subsample for whom $D = d^*$ and $M = m$ by $\hat{w}_4$ creates a pseudosample in which exposure to both d^* and m appears to have occurred at random, the mean of the outcome in this other pseudosample, which is given by $\sum {}^{I(D=d^*,M=m)\hat{w}_4 Y}/_{\sum I(D=d^*,M=m)\hat{w}_4}$, yields an estimate for $\mathbb{E}\left[Y\left(d^*, m\right)\right]$. Computing the difference between these weighted means following Equation 3.41 yields an estimate for the controlled direct effect.

Inverse probability weighting provides consistent estimates for the total, direct, and indirect effects of interest provided that the assumptions needed to identify these effects are all satisfied and provided that the models used to construct the weights are all correctly specified. If any identification assumptions are not met–for example, because there is unobserved confounding of the exposure-outcome or mediator-outcome relationships–then these estimators will suffer from systematic bias. Moreover, inverse probability weighting may also yield biased and inconsistent estimates if the GLMs used to construct the weights are incorrectly specified. If these models are misspecified, then the weights derived from them are also incorrect, and they will not transform the distribution of the sample data as intended, leading to systematic error.

Even when all identification assumptions are met and all models are correctly specified, however, the inverse probability weights defined previously may still yield imprecise and unstable estimates in finite samples (Cole and Hernan 2008; Robins et al. 2000; VanderWeele 2009b). This is because the weights involve inverse probabilities, which can be very large when the probabilities themselves are very small, and when some sample members are given very large weight, it may distort the effect estimates. The challenges stemming from extreme weights can be partly mitigated by using stabilized versions of the inverse probabilities, which involves scaling them down to have a mean equal to 1 and lower variance.

A set of stabilized weights for estimating total, natural direct, and natural indirect effects can be expressed as follows:

$$s\hat{w}_1 = \frac{\hat{P}\left(d^*\right)}{\hat{P}\left(d^*|C\right)},\ s\hat{w}_2 = \frac{\hat{P}\left(d\right)}{\hat{P}\left(d|C\right)},\ \text{and } s\hat{w}_3 = \frac{\hat{P}\left(d^*|C, M\right)\hat{P}\left(d\right)}{\hat{P}\left(d|C, M\right)\hat{P}\left(d^*|C\right)}. \tag{3.42}$$

In these expressions, $s\hat{w}_1 = \hat{w}_1 \times \hat{P}\left(d^*\right)$, that is, the unstabilized weight $\hat{w}_1$ has been scaled down by multiplying it with $\hat{P}\left(d^*\right)$, the predicted marginal probability of exposure to d^*. Similarly, $s\hat{w}_2 = \hat{w}_2 \times \hat{P}\left(d\right)$ and $s\hat{w}_3 = \hat{w}_3 \times \hat{P}\left(d\right)$, where both $\hat{w}_2$ and $\hat{w}_3$ have been scaled down by multiplying them with $\hat{P}\left(d\right)$, the predicted marginal probability of exposure to d.

A stabilized weight for estimating controlled direct effects is given by the following expression:

$$s\hat{w}_4 = \frac{\hat{P}\left(M|D\right)\hat{P}\left(D\right)}{\hat{P}\left(M|C, D\right)\hat{P}\left(D|C\right)}, \tag{3.43}$$

which comes from scaling down the unstabilized weight $\hat{w}_4$ by multiplying it with $\hat{P}\left(M|D\right)\hat{P}\left(D\right)$. The first term in this scaling factor, $\hat{P}(M|D)$, is the predicted probability that a sample member experiences their observed value of the mediator given their exposure, while the second term, $\hat{P}(D)$, is the predicted marginal probability that a sample member experiences their observed exposure.

Using stabilized rather than unstabilized weights in the estimation procedures delineated previously will yield less erratic and less variable effect estimates. Stabilized weights are therefore always preferred when estimating total, direct, and indirect effects. To construct estimates based on the stabilized weights, analysts need only compute $\{s\hat{w}_1, s\hat{w}_2, ..., s\hat{w}_4\}$ and then substitute these weights for $\{\hat{w}_1, \hat{w}_2, ..., \hat{w}_4\}$ in Equations 3.40 and 3.41 above.

The performance of weighting estimators can sometimes be improved even further by censoring the weights. Censoring the weights involves top and bottom coding very large and very small weights, respectively, to reduce to the influence of outliers, and by extension, to improve the precision of effect estimates. For example, researchers might compute effect estimates after censoring the weights at their 1st and 99th percentiles in the sample data. If less censoring is desired, the 0.1 and 99.9 percentiles can be used. Alternatively, if more censoring is desired, the 5th and 95th percentiles, or maybe even the 10th and 90th percentiles, would be appropriate. In general, the greater the degree of censoring, the more stable and less variable are the weights, and consequently, also the effect estimates based thereon. But this improved stability comes at the cost of greater systematic bias, as censoring the weights more and more essentially unwinds the distributional transformations that weighting is designed to achieve in the first place. For example, censoring $s\hat{w}_1$ and $s\hat{w}_2$ would partially unbalance the distribution of the confounders across levels of the exposure, thereby reintroducing some bias due to uncontrolled confounding. In many applications, however, a small amount of censoring can lead to large improvements in the stability of estimates without inducing nontrivial biases.

Table 3.4 reports the results of an analysis on the NLSY using inverse probability weighting. Specifically, in this table, we report point estimates for the total, direct, and indirect effects of college attendance on standardized CES-D scores, with unemployment status as the focal mediator. To estimate these effects, we implemented the weighting estimators using a logit model for $f(D|C)$, the conditional distribution of college attendance given the baseline confounders. We also used a logit model for $s(D|C, M)$, the conditional distribution of college attendance given the baseline confounders and unemployment status, as well as for $g(M|C, D)$, the conditional distribution of unemployment status given the confounders and college attendance. We fit these models to the NLSY by the method of maximum likelihood and then used them to construct a set of stabilized weights, as in Equations 3.42 and 3.43. Links to the code and data for implementing these analyses are provided in the table footnote.

Effect estimates based on inverse probability weighting are similar to those reported previously in Tables 3.2 and 3.3. Consistent with our prior results, the weighting estimates also provide little evidence that unemployment mediates the effect of college attendance on later life depression. The weighting estimate for the $ATE(1, 0)$ suggests that attending college would reduce depression measured later at age 40 by nearly one-fifth of a standard deviation, while estimates for both the $NDE(1, 0)$ and $CDE(1, 0, 0)$ are comparable to this effect. The weighting estimate for the $NIE(1, 0)$, by contrast, is close to zero, which suggests that unemployment does not play a very important mediating role in the causal chain connecting education to mental health. The consistent pattern of results given by estimators based on linear models, simulation, and weighting indicates that this conclusion is robust to the different modeling assumptions required of each approach.

To summarize, the weighting estimators described in this section offer an alternative approach to the methods presented in Sections 3.5.1 and 3.5.2. In contrast to methods based on linear models or simulation,

Table 3.4: Total, Direct, and Indirect Effects of College Attendance on CES-D Scores as Estimated from the NLSY using Inverse Probability Weighting.

Estimand	Point Estimates Additive Logit Models for $f(D\|C)$, $s(D\|C,M)$, and $g(M\|C,D)$
$ATE(1,0)$	$-.173$
$NDE(1,0)$	$-.166$
$NIE(1,0)$	$-.008$
$CDE(1,0,0)$	$-.146$

Note: Estimates are expressed in standard deviation units. D denotes college attendance, M unemployment status, and C the baseline confounders. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/table_3-4`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

inverse probability weighting only requires models for the exposure and mediator, rather than for the mediator and outcome. Although inverse probability weighting is flexible and can be implemented with a broad class of GLMs, it is generally best suited for applications where the exposure and mediator are discrete and take on relatively few values. In applications where these variables take on many values or are continuous, inverse probability weighting may perform poorly due to unreliable estimates of the conditional probabilities or conditional densities needed to construct the weights. In this situation, seemingly minor misspecification errors can lead to large distortions in the weights, which in turn can lead to large biases in effect estimates. When the exposure and mediator are binary, ordinal, or polytomous, and the researcher is confident in their ability to model the distribution of these variables, inverse probability weighting may be preferred over other options. Otherwise, estimators based on linear models or simulation may provide more accurate results. As a general practice, researchers should experiment with multiple estimation strategies to assess the robustness of their findings to different modeling assumptions.

3.6 Statistical Inference

In Sections 3.4 and 3.5, we concentrated on point estimation of total, direct, and indirect effects, which involves assigning a single numeric value to a given estimand using data from a random sample. However, this point estimate may deviate from the true estimand due to sampling error or systematic bias. Provided that the assumptions needed to identify an estimand from observable data are satisfied and any models used for estimation are correctly specified, point estimates from the methods discussed previously will not suffer from any significant bias in sufficiently large samples, as all the estimators we have presented so far are consistent. Nevertheless, due to the random variability that arises from sampling, these estimates will still differ from their target estimands. In this section, we illustrate how to quantify the uncertainty that afflicts our point estimates because they are based on data from a random sample, rather than the entire population of interest.

Recall that every estimator is a function of a random sample and, therefore, has a probability distribution. This distribution is called the sampling distribution, and it describes the range of possible estimates and their associated probabilities if an estimator were applied repeatedly to all possible samples of size n from the target population. In other words, the sampling distribution is the probability distribution of estimates generated by the repeated application of an estimator to random samples from a target population. The shape and spread of this distribution reflect the degree of uncertainty, or sampling variability, affecting an estimator. Understanding the sampling distribution is therefore crucial for quantifying the precision of an estimator and drawing inferences about its estimand.

To quantify this precision and draw inferences, we use a method called the nonparametric bootstrap (Davison and Hinkley 1997; Efron and Tibshirani 1994), which constructs an approximate sampling distribution using data from our original sample of the target population. The nonparametric bootstrap belongs to a class of procedures known as re-sampling methods, and it is a versatile approach that can be applied with nearly all of the estimators covered in this book. It is especially useful when the sampling distribution of an estimator is either unknown or difficult to derive analytically. Although some estimators have known sampling distributions that can be analytically derived, others do not. Furthermore, the sampling distribution of an estimator may only be known or derived in certain settings or with additional assumptions that may not be valid in many social science applications. We therefore adopt the nonparametric bootstrap throughout this book, as it can provide valid inferential statistics for total, direct, and indirect effects under general conditions and with a broad class of estimators.

In substantive terms, the nonparametric bootstrap treats the original sample data as a representation of the larger population of interest. To construct a new "bootstrap sample," individuals included in the original data are re-sampled with replacement. Then, using this bootstrap sample, some estimate or set of estimates, such as the $\widehat{NDE}(d, d^*)^{ipw}$ or $\widehat{NIE}(d, d^*)^{ipw}$, are computed following the same procedures with which the original data were analyzed. This process is repeated many times, with new bootstrap samples constructed at each step via sampling with replacement from the original data. After many repetitions, a representation of the sampling distribution can be constructed empirically using the full set of bootstrap estimates. This empirical representation of the sampling distribution, known as the *bootstrap distribution*, can be used to quantify the degree of random variability in an estimator and to draw inferences about the target estimand. Specifically, we focus on how to use the bootstrap distribution to construct confidence intervals and to conduct null hypothesis tests.

The objective when constructing a τ-percent confidence interval for some target estimand θ, such as an average total effect or a natural direct effect, is to provide a range of estimates that contain the true but unknown value for the target estimand τ percent of the time under repeated sampling. This means that if it were possible to draw independent samples from a target population repeatedly, and to compute a confidence interval for each of these samples, then τ percent of the intervals would contain the true value of the target estimand, while $100 - \tau$ percent would not cover its true value. Thus, confidence intervals quantify the uncertainty in an analysis due to random variability by providing a range of estimates that cover the target estimand with known probability in repeated sampling. For a given level of τ, the wider the confidence interval, the greater is the degree of uncertainty due to sampling error.

A τ-percent confidence interval for a target estimand θ can be constructed using the nonparametric bootstrap as follows:

1. **Randomly sample the observed data with replacement.** Select a sample of size n from the observed data at random, where n is the number of original sample members and where each member drawn from the original sample is returned to the sampling frame after selection. This new sample assembled by re-sampling the observed data with replacement is called a bootstrap sample.

2. **Compute an estimate.** Use the bootstrap sample from step 1 to compute a point estimate for the

target estimand θ. This estimate, denoted by $\hat{\theta}_b$, is called a bootstrap estimate.

3. **Repeat the prior steps many times.** For $b = 1, 2, ..., B$, repeat steps 1 and 2 from above, saving the value of $\hat{\theta}_b$ from each iteration. In most applications, $10^3 \leq B \leq 10^4$ will provide a sufficiently accurate representation of the sampling distribution.

4. **Compute percentiles.** The lower bound of a τ-percent confidence interval is given by the $0.5\,(100 - \tau)$th percentile of the bootstrap estimates, $\left\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B\right\}$, while the upper bound is given by the $(100 - 0.5\,(100 - \tau))$th percentile of these estimates.

To illustrate this procedure more concretely, consider the following example. Suppose we want to construct a $\tau = 95$ percent confidence interval for $\theta = NIE\,(d, d^*)$, which we intend to estimate using the method of inverse probability weighting. How would we construct this confidence interval using the nonparametric bootstrap?

First, we would draw a sample of size n from the observed data with replacement. To this end, we would randomly select one individual from the observed data and then immediately return them to the original sample. Next, we would randomly select another individual from the observed data and then return them to the original sample as well. We would repeat this sampling process over and over until a total of n individuals have been selected for the bootstrap sample. Because individuals are returned to the original sample upon their selection, the bootstrap sample will typically contain replicate cases–that is, members of the original sample may be selected and included in a bootstrap sample multiple times.

Second, we would compute an estimate for the target estimand using this bootstrap sample. Specifically, because we elected to use inverse probability weighting and are targeting the natural indirect effect, we would compute $\widehat{NIE}\,(d, d^*)^{ipw}$ exactly as described in Section 3.5.3, only now we would perform these calculations on the bootstrap sample. We would then save the resulting estimate, denoted by $\hat{\theta}_1$, for later use.

Third, to obtain a full set of bootstrap estimates, we would repeat the previous steps a large number of times, say $B = 10^3$. This would yield a collection of bootstrap estimates denoted by $\left\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_{1000}\right\}$. Because each bootstrap estimate is computed using a different sample from the observed data, they may differ from one another, but taken together, their empirical distribution provides an approximation for the sampling distribution of our estimator. In other words, the distribution of the bootstrap estimates, assembled by repeatedly sampling from the observed data with replacement, approximates the sampling distribution of $\widehat{NIE}\,(d, d^*)^{ipw}$ that we would obtain were it possible to compute these estimates on repeated samples drawn from the target population of interest.

Finally, we would compute the upper and lower bounds of our desired confidence interval by finding the $0.5\,(100 - 95) = 2.5$th and $(100 - 0.5\,(100 - 95)) = 97.5$th percentiles, respectively, of the bootstrap estimates, where the kth percentile is the value at or below which k percent of all bootstrap estimates fall. For example, if we were to rank all 1000 bootstrap estimates, $\left\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_{1000}\right\}$, from lowest to highest, the 2.5th percentile would be the value of the estimate with rank 25, while the 97.5th percentile would be the value of the estimate with rank 975. These bounds together would form our 95 percent confidence interval for the $NIE\,(d, d^*)$. Assuming this estimand is identified and all models used to implement inverse probability weighting are correctly specified, the interval provides a range of effect estimates that contain the true but unknown value of the natural indirect effect with 95 percent probability under repeated sampling from the target population.

Another approach to drawing inferences in the presence of sampling uncertainty is to perform a null hypothesis test. In such a test, the analyst aims to determine whether or not a target estimand is equal to a specific value in the population of interest, using the information contained in the sample data. The test begins with the analyst stating a null hypothesis about an estimand. For example, they might hypothesize that the natural indirect effect of D on Y via M is zero in the target population. The objective is then to assess the extent to which the sample data are inconsistent with this hypothesis. However, because of sampling error, the point estimate for the target estimand obtained from the sample data may differ from its hypothesized value, even when the null hypothesis is true. Thus, to evaluate whether the observed data are inconsistent with the null hypothesis, we need to use probabilistic methods that account for random variability in our point estimates due to sampling.

The degree of evidence that the sample data provide against a null hypothesis can be summarized using a p-value (Wasserstein and Lazar 2016). A p-value is the probability of observing a point estimate as or even more extreme than the one obtained from the sample data, assuming the null hypothesis is true and repeated samples are drawn from the target population. If this probability is very small, it indicates that obtaining the point estimate we computed from the observed data is unlikely given the null hypothesis. We would therefore conclude that the sample data are inconsistent with the null hypothesis and reject the possibility of its truth. Conversely, if the p-value is not very small, we would fail to reject the null hypothesis and conclude that the estimate given by the sample data is not inconsistent with it. This does not necessarily mean that the null hypothesis is true, but rather that the sample data do not provide enough evidence to reject it.

To test the null hypothesis that a target estimand θ is equal to some fixed value t, we can compute a p-value by inverting a bootstrap confidence interval. This approach relies on the fact that the p-value for such a test is the smallest value of $(100 - \tau)/100$ such that t is not contained in the corresponding τ-percent confidence interval. With a set of bootstrap estimates $\left\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B\right\}$ obtained using the methods outlined previously, we can compute this p-value as $2 \times \min\left(\hat{\pi}\left(\hat{\theta}_b < t\right), \hat{\pi}\left(\hat{\theta}_b > t\right)\right)$, where $\hat{\pi}\left(\hat{\theta}_b < t\right)$ and $\hat{\pi}\left(\hat{\theta}_b > t\right)$ respectively denote the proportion of bootstrap estimates that are less than and greater than the hypothesized value for the target parameter. This quantity approximates the probability of observing a point estimate as extreme or more extreme than the one given by our sample data, if the target estimand θ were in fact equal to t. In general, the smaller the p-value, the more inconsistent are the sample data with the null hypothesis. Conventionally, p-values less than .05 are considered sufficiently inconsistent with the null hypothesis to reject the possibility that it may be true, although this criterion sometimes differs between disciplines and sub-fields.

Although the nonparametric bootstrap is our recommended approach to statistical inference, it is not without limitations. The method is computationally intensive, as it requires repeatedly sampling from the observed data and computing new estimates thousands of times. While increases in computational power and the availability of parallel processing have made bootstrapping more tractable, there are still some applications where it can be time-consuming. The nonparametric bootstrap is also more difficult to use with data from complex sample designs, such as those involving stratification, clustering, or multistage selection. To account for these design features, the procedure can be modified, for example, by re-sampling clusters of observations from within strata, but this requires additional care to ensure the method appropriately reflects the sampling process used to collect the original data. In such cases, researchers can consult the technical literature on bootstrapping (e.g., Davison and Hinkley 1997; Efron and Tibshirani 1994; Rao and Wu 1988) for guidance about the most appropriate implementation given the details of their sample design.

Table 3.5 presents a set of inferential statistics from our analysis of the NLSY using the nonparametric bootstrap. Specifically, the first column of the table contains 95% confidence intervals for the total, direct,

Table 3.5: Inferential Statistics for Total, Direct, and Indirect Effects of College Attendance on CES-D Scores Computed from the NLSY using Inverse Probability Weighting and the Nonparametric Bootstrap.

Estimands	95% Bootstrap Confidence Interval	Bootstrap P-value for Null of No Effect
$ATE(1,0)$	$[-.274, -.063]$	.003
$NDE(1,0)$	$[-.267, -.056]$	.005
$NIE(1,0)$	$[-.027,\ .004]$	.215
$CDE(1,0,0)$	$[-.242, -.039]$	.009

Note: Estimates are based on $B = 2000$ bootstrap replications. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/table_3-5`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

and indirect effects of college attendance on depression, as mediated by unemployment. They were computed by drawing $B = 2000$ bootstrap samples from the NLSY, computing effect estimates using the method of inverse probability weighting on each of these samples, and then by finding the 2.5th and 97.5th percentiles of the bootstrap estimates. The resulting intervals provide a range of estimates that contain the true values of the target estimands with 95 percent probability under repeated sampling, assuming that these effects are identified and there is no model misspecification. In general, our interval estimates suggest that the total and direct effects of college attendance are negative but are also estimated rather imprecisely, as indicated by the relatively wide range of values contained in the intervals for the $ATE\,(0,1)$, $NDE\,(1,0)$, and $CDE\,(1,0,0)$. The confidence interval for the $NIE\,(1,0)$, on the other hand, is not as wide, and it spans a range of effect estimates that are all substantively small.

To illustrate how these intervals were constructed, Figure 3.7 displays a histogram of the $B = 2000$ bootstrap estimates for the natural indirect effect. The distribution of bootstrap estimates has a slight negative skew and is slightly more peaked than a normal distribution, which is bell-shaped and perfectly symmetric. The dashed vertical lines denote the 2.5th and 97.5th percentiles of this distribution, and thus they represent the lower and upper bounds, respectively, of the 95% confidence interval.

The second column of Table 3.5 presents p-values from tests of the null hypothesis that a target estimand is equal to zero. These p-values are also based on the nonparametric bootstrap and estimates obtained via inverse probability weighting. Specifically, they were computed by constructing bootstrap distributions of estimates and then by finding the smallest value of $(100 - \tau)/100$ such that the hypothesized value of zero for each target estimand is not contained in the corresponding τ-percent confidence interval.

The p-value from a test of the null hypothesis that $ATE\,(1,0) = 0$ is .003, which indicates that the NLSY provides considerable evidence against the possibility that college attendance has no overall effect on depression. Similarly, tests of whether the $NDE\,(1,0)$ and $CDE\,(1,0,0)$ are equal to zero both yield small p-values as well. These results indicate that data from the NLSY are also inconsistent with the possibility that there is no natural or controlled direct effect of college attendance on depression. By contrast, a test of the null hypothesis that the $NIE\,(1,0)$ is equal to zero yields a p-value of .215. This result indicates that the data are not inconsistent with the possibility that there is no indirect effect of college attendance on depression operating through unemployment.

To illustrate how these p-values were calculated, Figure 3.8 displays a kernel density plot for the $B = 2000$ bootstrap estimates of the natural indirect effect. The darkly shaded area represents

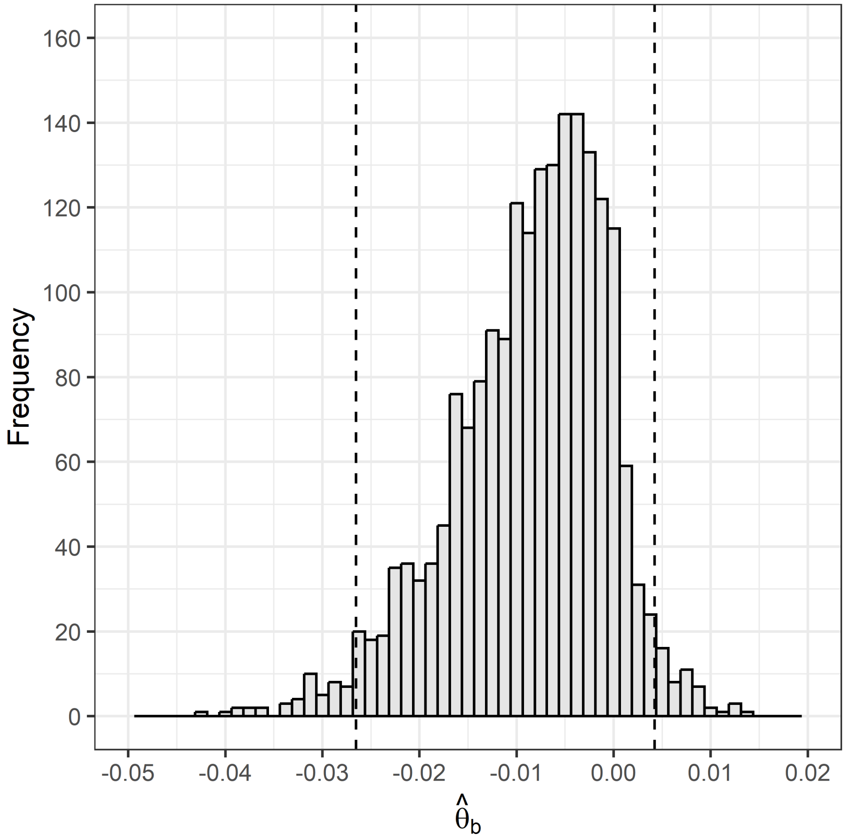


Figure 3.7: Histogram of Bootstrap Estimates for $\widehat{NIE}(1,0)^{ipw}$ based on the NLSY.

Note: Distribution is based on $B = 2000$ bootstrap replications. Dashed vertical lines denote the 2.5th and 97.5th percentiles of the distribution and thus represent the lower and upper bounds, respectively, of a 95% bootstrap confidence interval. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/figure_3-7`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

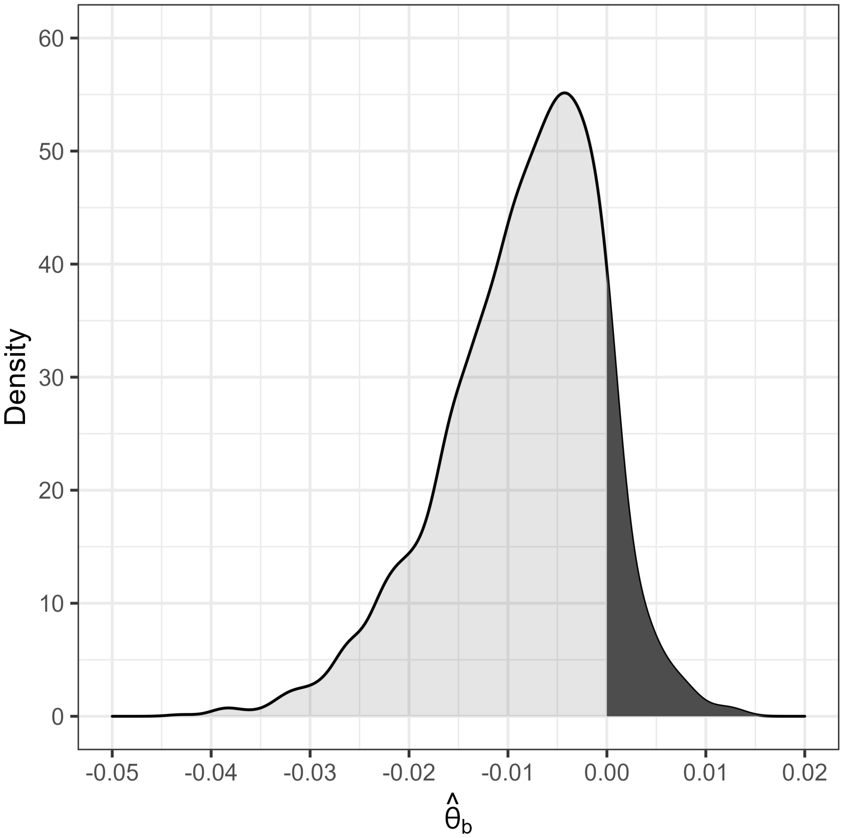


Figure 3.8: Kernel Density Plot of Bootstrap Estimates for $\widehat{NIE}(1,0)^{ipw}$ based on the NLSY.

Note: Distribution is based on $B = 2000$ bootstrap replications. The darkly shaded area represents $\min\left(\hat{\pi}\left(\hat{\theta}_b < 0\right), \hat{\pi}\left(\hat{\theta}_b > 0\right)\right)$, which, in this case, is the proportion of bootstrap estimates greater than zero. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/figure_3-8`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

$\min\left(\hat{\pi}\left(\hat{\theta}_b < 0\right), \hat{\pi}\left(\hat{\theta}_b > 0\right)\right)$, which is $\hat{\pi}\left(\hat{\theta}_b > 0\right) = .1075$ in this instance. Twice this proportion, $.1075 \times 2 = .215$, is the p-value given by inverting a percentile bootstrap interval. That is, if we were to construct a τ-percent confidence interval by finding the $0.5\,(100 - \tau)$th and $(100 - 0.5\,(100 - \tau))$th percentiles of the bootstrap estimates, the interval corresponding to $\tau = 78.5$ gives the smallest value for $(100 - \tau)/100$ while still excluding the hypothesized value of zero for the target estimand. By extension, a value of τ any larger than 78.5 yields a confidence interval that covers zero based on the corresponding percentiles of the bootstrap estimates, and a value smaller than 78.5 excludes zero from the corresponding interval but fails to minimize $(100 - \tau)/100$. Thus, the p-value from our test of the null hypothesis that $NIE\,(1, 0) = 0$ is $(100 - 78.5)/100 = .215$, or twice the proportion of bootstrap estimates that are greater than the value hypothesized for the indirect effect under the null.

3.7 Sensitivity Analysis

An important concern in analyses of causal mediation is the problem of unobserved confounding. If there are unobserved confounders for the exposure-outcome, mediator-outcome, or exposure-mediator relationships, then all of the estimators that we discussed previously would be biased, and they would not converge to their target estimands with increases in sample size. Consequently, our statistical inferences would be *invalid*: that is, with a biased and inconsistent estimator, τ-percent confidence intervals will not contain the true value of the target estimand τ percent of the time in repeated sampling, and p-values will not accurately reflect the probability of observing a point estimate as or more extreme than the one derived from the observed data under the null hypothesis. Thus, in the presence of unobserved confounding, our conclusions about causal mediation can be mistaken, regardless of how much data we collect.

Confounding bias poses a significant challenge in analyses of causal mediation because it is exceedingly difficult to conduct experiments that ensure the relationships among key variables are all unconfounded by design. When confounding cannot be controlled by experimental design, formal sensitivity analyses are useful for assessing potential biases and their impact on our inferences. In this section, we introduce a set of bias formulas for total, direct, and indirect effects in the presence of unobserved confounding (VanderWeele 2010, 2015; VanderWeele and Arah 2011). These formulas offer a means to evaluate the sensitivity of estimates to hypothetical patterns of unobserved confounding and to explore how the resulting bias may alter our conclusions about causal mediation.

We first present a set of bias formulas that possess the highest degree of generality. These formulas do not presuppose any particular technique or model for obtaining initial estimates of total, direct, and indirect effects. Moreover, they apply regardless of whether the exposure, mediator, or outcome is binary, ordinal, or continuous, and they place few restrictions on the pattern of unobserved confounding, making them compatible for use with any of the estimators discussed in this chapter. Nevertheless, there is a trade-off between generality and complexity. While the general formulas depend on few assumptions, they involve a large number of parameters, which can become unwieldy in practical applications. Thus, after introducing the bias formulas in their most general form, we explore additional simplifying assumptions that allow us to express them as straightforward functions of a few basic sensitivity parameters. This simplified approach is easier to use in practice because it requires specifying only a handful of parameters, but it hinges on stronger assumptions about the nature of unobserved confounding.

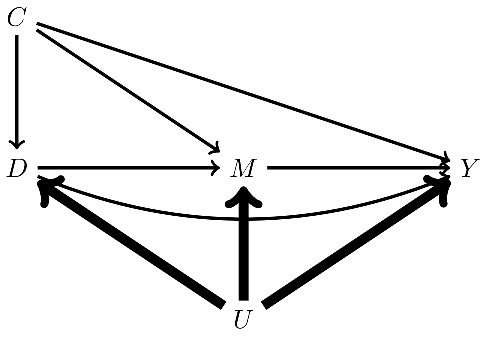


Figure 3.9: Graphical Illustration of Unobserved Exposure-outcome, Mediator-outcome, and Exposure-mediator Confounding.

Note: D denotes the exposure, M denotes a mediator, Y denotes the outcome, C denotes a set of observed baseline confounders, and U denotes an unobserved confounder.

3.7.1 Nonparametric Bias Formulas

Consider a scenario in which an unobserved variable, denoted by U, affects the exposure D, the mediator M, and the outcome Y, as depicted in Figure 3.9. Because the unobserved variable confounds the exposure-outcome, mediator-outcome, and exposure-mediator relationships, all of the conditional independence assumptions required to identify total, direct, and indirect effects are violated, and by extension, all of the estimators we have considered thus far would be biased and inconsistent. In this situation, the bias afflicting an estimator for the natural direct effect is given by

$$\begin{aligned}\text{Bias}\left(\widehat{NDE}\left(d,d^*\right)\right) &= \sum_{m,u,c}\mathbb{E}\left[Y|c,d,m,u\right]P\left(m|c,d^*\right)P\left(c\right)\left(P\left(u|c,d,m\right)-\frac{P\left(u|c,d^*,m\right)P\left(u|c\right)}{P\left(u|c,d^*\right)}\right)\\ &\quad-\sum_{m,u,c}\mathbb{E}\left[Y|c,d^*,m,u\right]P\left(m|c,d^*\right)P\left(c\right)\\ &\quad\times\left(P\left(u|c,d^*,m\right)-\frac{P\left(u|c,d^*,m\right)P\left(u|c\right)}{P\left(u|c,d^*\right)}\right), \qquad (3.44)\end{aligned}$$

the bias afflicting an estimator for the natural indirect effect is given by

$$\begin{aligned}\text{Bias}\left(\widehat{NIE}\left(d,d^*\right)\right) &= \sum_{m,u,c}\mathbb{E}\left[Y|c,d,m,u\right]P\left(m|c,d\right)P\left(c\right)\left(P\left(u|c,d,m\right)-\frac{P\left(u|c,d,m\right)P\left(u|c\right)}{P\left(u|c,d\right)}\right)\\ &\quad-\sum_{m,u,c}\mathbb{E}\left[Y|c,d,m,u\right]P\left(m|c,d^*\right)P\left(c\right)\\ &\quad\times\left(P\left(u|c,d,m\right)-\frac{P\left(u|c,d^*,m\right)P\left(u|c\right)}{P\left(u|c,d^*\right)}\right), \qquad (3.45)\end{aligned}$$

and the bias afflicting an estimator for the total effect is given by the sum of these two expressions, $\text{Bias}\left(\widehat{ATE}\left(d,d^*\right)\right) = \text{Bias}\left(\widehat{NDE}\left(d,d^*\right)\right) + \text{Bias}\left(\widehat{NIE}\left(d,d^*\right)\right)$. Similarly, the bias in an estimator for the controlled direct effect is given by

$$\begin{aligned}\text{Bias}\left(\widehat{CDE}\left(d,d^{*},m\right)\right) =& \sum_{u,c}\mathbb{E}\left[Y|c,d,m,u\right]P\left(c\right)\left(P\left(u|c,d,m\right)-\frac{P\left(u|c,d,m\right)P\left(u|c\right)}{P\left(u|c,d\right)}\right)\\ &-\sum_{u,c}\mathbb{E}\left[Y|c,d^{*},m,u\right]P\left(c\right)\\ &\times\left(P\left(u|c,d^{*},m\right)-\frac{P\left(u|c,d^{*},m\right)P\left(u|c\right)}{P\left(u|c,d^{*}\right)}\right).\end{aligned} \tag{3.46}$$

The expressions presented above are highly complex. In general, they reveal that the biases affecting estimators for direct and indirect effects depend on how the outcome differs across levels of the unobserved confounder, given the exposure, mediator, and baseline confounders. Moreover, they also reveal that the biases depend on how the unobserved confounder differs across levels of the exposure and mediator. Specifying plausible values for all the different quantities that compose these bias terms is exceptionally challenging, as this would require a level of prior subject-matter knowledge that is almost never available in the social sciences. Consequently, it is usually necessary to introduce simplifying assumptions in order to facilitate the practical application of these bias expressions in a sensitivity analysis.

3.7.2 Bias from Exposure-outcome Confounding

To this end, suppose that the unobserved variable U only confounds the relationship between the exposure D and the outcome Y. Moreover, suppose further that U is binary and that $\mathbb{E}\left[Y|c,d,m,U=1\right]-\mathbb{E}\left[Y|c,d,m,U=0\right]$ is constant across levels of C, D, and M, or in other words, assume that the difference in the mean of the outcome across levels of the unobserved confounder does not vary with the exposure, mediator, or baseline confounders. Finally, suppose that $P\left(U=1|c,d\right)-P\left(U=1|c,d^{*}\right)$ is constant across levels of C, or in other words, assume that the difference in the distribution of the unobserved confounder comparing exposure d versus d^{*} does not vary with the baseline confounders. Under these simplifying assumptions, an estimator for the natural indirect effect that does not adjust for U will remain unbiased. Conversely, without adjusting for U, estimators for the natural and controlled direct effects will be biased.

In this scenario, the biases afflicting estimators for the natural and controlled direct effects are equal, and they can both be expressed as follows:

$$\text{Bias}\left(\widehat{NDE}\left(d,d^{*}\right)\right)=\text{Bias}\left(\widehat{CDE}\left(d,d^{*},m\right)\right)=\delta_{UY|C,D,M}\times\delta_{DU|C}, \tag{3.47}$$

where $\delta_{UY|C,D,M}=\mathbb{E}\left[Y|c,d,m,U=1\right]-\mathbb{E}\left[Y|c,d,m,U=0\right]$ and $\delta_{DU|C}=P\left(U=1|c,d\right)-P\left(U=1|c,d^{*}\right)$. This simple expression is a function of two sensitivity parameters, $\delta_{UY|C,D,M}$ and $\delta_{DU|C}$, which are easier to specify with plausible values in practice. The first sensitivity parameter $\delta_{UY|C,D,M}$ captures the difference in the mean of the outcome associated with a unit increase in the unobserved confounder, conditional on the baseline confounders and mediator, while the second sensitivity parameter $\delta_{DU|C}$ captures the difference in the probability of the unobserved confounder comparing level d versus d^{*} of the exposure, conditional on the confounders. Thus, under several simplifying assumptions, the bias in estimators for both natural and controlled direct effects is equal to a "partial effect" of the unobserved confounder on the outcome multiplied by a "partial effect" of the exposure on the unobserved confounder.

3.7.3 Bias from Mediator-outcome Confounding

Consider next the scenario in which an unobserved variable U only affects the mediator M and the outcome Y, but not the exposure D. This pattern of unobserved confounding may arise in an experimental study where the exposure, but not the mediator, is randomly assigned. Random assignment of the exposure would ensure that there is no unobserved confounding of the exposure-outcome and exposure-mediator relationships. However, it would not obviate the problem of mediator-outcome confounding by unobserved factors. In this scenario, estimators for direct and indirect effects would suffer from bias, while estimators for the total effect would remain unbiased.

Specifically, under the same set of simplifying assumptions outlined previously in Section 3.7.2, the bias in estimators for natural and controlled direct effects arising from mediator-outcome confounding are again equal, and they can both be expressed as follows:

$$\text{Bias}\left(\widehat{NDE}\left(d,d^*\right)\right)=\text{Bias}\left(\widehat{CDE}\left(d,d^*,m\right)\right)=\delta_{UY|C,D,M}\times\delta_{DU|C,M},\tag{3.48}$$

where $\delta_{UY|C,D,M} = \mathbb{E}\left[Y|c,d,m,U=1\right] - \mathbb{E}\left[Y|c,d,m,U=0\right]$ and $\delta_{DU|C,M} = P\left(U=1|c,d,m\right) - P\left(U=1|c,d^*,m\right)$. In this expression, $\delta_{UY|C,D,M}$ is defined exactly as in Section 3.7.2 above, while $\delta_{DU|C,M}$ represents the difference in the probability of the unobserved confounder comparing level d versus d^* of the exposure, now conditional on both the baseline confounders and the mediator.

Moreover, when U only confounds the $M \rightarrow Y$ relationship, the bias in an estimator for the natural indirect effect is now equal to the negation of the bias in the direct effects, under the simplifying assumptions outlined previously. Specifically, the bias in this scenario can be expressed as follows:

$$\text{Bias}\left(\widehat{NIE}\left(d,d^*\right)\right)=-\delta_{UY|C,D,M}\times\delta_{DU|C,M},\tag{3.49}$$

where both sensitivity parameters, $\delta_{UY|C,D,M}$ and $\delta_{DU|C,M}$, are defined exactly as before.

The form of Equations 3.48 and 3.49 demonstrates that unobserved mediator-outcome confounding distorts how our estimators partition the total effect into direct and indirect components but does not lead to systematic error in estimates of the total effect itself. This is because the sum of these two expressions gives the bias in an estimator for the total effect that does not adjust for U, which in this case is equal to zero.

3.7.4 Bias from Exposure-mediator Confounding

Now consider the scenario in which an unobserved variable U confounds only the relationship between the exposure D and mediator M. When U only confounds the $D \rightarrow M$ relationship, an estimator for the controlled direct effect that does not adjust for U will remain unbiased. Under the assumption that the controlled direct effect does not depend on m within levels of the baseline confounders, an estimator for the natural direct effect that does not adjust for U will also be unbiased in this situation. Conversely, without adjusting for U, an estimator for the natural indirect effect will generally suffer from confounding bias. If U is binary, $P\left(U=1|c,d\right)-P\left(U=1|c,d^*\right)$ is constant across levels of C, and $P\left(m|c,d,U=1\right)-P\left(m|c,d,U=0\right)$ is constant across levels of D, the bias in an estimator for the natural indirect effect is given by the following expression:

$$\text{Bias}\left(\widehat{NIE}\left(d,d^*\right)\right)=\delta_{DU|C}\times\delta_{UMY|C}.\tag{3.50}$$

In this bias formula, $\delta_{DU|C}$ is equal to $P(U=1|c,d) - P(U=1|c,d^*)$, as in Section 3.7.2. It represents the difference in the probability of the unobserved confounder comparing level d versus d^* of the exposure, conditional on the baseline confounders. The other sensitivity parameter in this expression, $\delta_{UMY|C}$, is equal to $\sum_{m,c} \mathbb{E}[Y|c,d,m](P(m|c,d,U=1) - P(m|c,d,U=0))P(c)$. It captures how the unobserved confounder U influences the outcome Y through its effect on the mediator M. In other words, it is similar to an "indirect effect" of the unobserved confounder on the outcome that operates through the mediator.

3.7.5 Bias-adjusted Effect Estimates

Table 3.6 summarizes the bias formulas for each of our target estimands under the simplifying assumptions outlined previously. With these bias formulas, a formal sensitivity analysis proceeds by reevaluating the focal effect estimates across different hypothetical patterns of unobserved confounding. We would begin by specifying the bias formulas with plausible values for their sensitivity parameters, and then we would construct a set of bias-adjusted effect estimates by subtracting the bias terms from their corresponding point estimates.

Specifically, a set of bias-adjusted estimates for the total, direct, and indirect effects of interest can be expressed as follows:

$$\begin{aligned}
\widehat{NDE}(d,d^*)^{adj} &= \widehat{NDE}(d,d^*) - \text{Bias}\left(\widehat{NDE}(d,d^*)\right) \\
\widehat{NIE}(d,d^*)^{adj} &= \widehat{NIE}(d,d^*) - \text{Bias}\left(\widehat{NIE}(d,d^*)\right) \\
\widehat{ATE}(d,d^*)^{adj} &= \widehat{ATE}(d,d^*) - \text{Bias}\left(\widehat{ATE}(d,d^*)\right) \\
\widehat{CDE}(d,d^*,m)^{adj} &= \widehat{CDE}(d,d^*,m) - \text{Bias}\left(\widehat{CDE}(d,d^*,m)\right),
\end{aligned} \tag{3.51}$$

where $\widehat{NDE}(d,d^*)$, $\widehat{NIE}(d,d^*)$, $\widehat{ATE}(d,d^*)$ and $\widehat{CDE}(d,d^*,m)$ each denote an estimator from Sections 3.4 to 3.5, and the "adj" superscript indicates that they have been adjusted for bias due to an assumed pattern of unobserved confounding. The degree to which our inferences are sensitive to unobserved confounding can be assessed by evaluating the bias-adjusted estimates across a range of plausible values for the sensitivity parameters and examining whether the adjusted estimates depart from hypothesized patterns. Confidence intervals for the bias-adjusted estimates can be constructed using the nonparametric bootstrap.

3.8 An Empirical Illustration: The Effect of Job Training on Employment

In this section, we illustrate the methods described previously with a reanalysis of data from the JOBSII study (Vinokur et al. 1995; Vinokur and Schul 1997; Imai et al. 2010a). The JOBSII study was a randomized field experiment designed to investigate the effect of participating in a job training workshop on subsequent employment among a set of unemployed workers. The experiment randomly assigned $n = 1801$ unemployed workers to treatment and control groups. In the treatment group, participants were enrolled in a workshop where they learned about job search skills and strategies for coping with setbacks. In the control group, participants received only a booklet providing basic job search advice. The researchers who conducted this experiment hypothesized that workshop participation would increase reemployment in part by enhancing participants' confidence in their ability to search for a job, and they collected data on what they called "job

Table 3.6: Simplified Bias Formulas for Total, Direct, and Indirect Effects.

Bias/estimator	Type of Confounding		
	$D \leftarrow U \rightarrow Y$	$M \leftarrow U \rightarrow Y$	$D \leftarrow U \rightarrow M$
Bias $\left(\widehat{CDE}\left(d, d^{*}, m\right)\right)$	$\delta_{UY\mid C,D,M} \times \delta_{DU\mid C}$	$\delta_{UY\mid C,D,M} \times \delta_{DU\mid C,M}$	0
Bias $\left(\widehat{NDE}\left(d, d^{*}\right)\right)$	$\delta_{UY\mid C,D,M} \times \delta_{DU\mid C}$	$\delta_{UY\mid C,D,M} \times \delta_{DU\mid C,M}$	0
Bias $\left(\widehat{NIE}\left(d, d^{*}\right)\right)$	0	$-\delta_{UY\mid C,D,M} \times \delta_{DU\mid C,M}$	$\delta_{DU\mid C} \times \delta_{UMY\mid C}$
Bias $\left(\widehat{ATE}\left(d, d^{*}\right)\right)$	$\delta_{UY\mid C,D,M} \times \delta_{DU\mid C}$	0	$\delta_{DU\mid C} \times \delta_{UMY\mid C}$

Note: These bias formulas variously assume that U is binary; $P\left(U=1 \mid c, d\right)-P\left(U=1 \mid c, d^{*}\right)$ is constant across levels of C; $\mathbb{E}\left[Y \mid c, d, m, U=1\right]-\mathbb{E}\left[Y \mid c, d, m, U=0\right]$ is constant across levels of C, D, and M; $P\left(m \mid c, d, U=1\right)-P\left(m \mid c, d, U=0\right)$ is constant across levels of the exposure D; and for the bias in natural direct effects under exposure-mediator confounding specifically, that the controlled direct effect does not depend on m within levels of the baseline confounders.

search self-efficacy" in a series of follow-up interviews. With these data, we examine whether job training increases reemployment and whether this effect is mediated by job search self-efficacy.

The exposure D is a binary variable coded 1 for those assigned to the job training workshop and 0 for those in the control group. The outcome Y is another binary variable coded 1 if a participant was working at least 20 hours per week following the workshop, and 0 otherwise. The mediator M is a multi-item index measuring job search self-efficacy. This index was constructed using six survey items that asked participants to rate, using a 5-point scale, their confidence in being able to successfully perform job search activities, such as completing a job application, leveraging their social network to discover job openings, and communicating effectively during a job interview. We also adjust for a set of baseline covariates, collectively denoted by C, measured before the workshop was conducted. They include measures of education, income, race, age, gender, and financial strain.

With these data, we focus on the following estimands. First, we examine the average total effect of job training on reemployment, which can be expressed as $ATE\left(1,0\right)=\mathbb{E}\left[Y\left(1\right)-Y\left(0\right)\right]$. Because the outcome Y is binary, this effect is equal to the difference in the probability of reemployment if individuals were to participate, versus not participate, in the job training workshop. Second, we examine the natural direct effect, which can be expressed as $NDE\left(1,0\right)=\mathbb{E}\left[Y\left(1, M\left(0\right)\right)-Y\left(0, M\left(0\right)\right)\right]$. This effect represents the difference in the probability of reemployment if individuals had, versus had not, attended the job training workshop and if they had experienced the level of self-efficacy that would have occurred for them had they not attended the workshop. It captures an effect of job training on employment that is not due to differences in self-efficacy induced by attending the workshop. Third, we focus on the natural indirect effect, which can be expressed as $NIE\left(1,0\right)=\mathbb{E}\left[Y\left(1, M\left(1\right)\right)-Y\left(1, M\left(0\right)\right)\right]$. This effect represents the difference in the probability of reemployment if individuals had attended the job training workshop and then experienced the level of self-efficacy that would have arisen in them had they attended rather than not attended the workshop. It captures an effect of job training on reemployment that operates specifically through a mechanism involving self-efficacy. Finally, we examine a controlled direct effect, which can be expressed as $CDE\left(1,0,4\right)=\mathbb{E}\left[Y\left(1,4\right)-Y\left(0,4\right)\right]$. This captures the effect of job training on the probability of reemployment if all individuals felt a relatively high level of self-efficacy, as indicated by $M=4$ on an index ranging from 1 to 5.

Table 3.7: Estimates of the Total, Direct, and Indirect Effects of Job Training on Employment from the JOBSII Study.

Estimands	Linear Model Estimates		Simulation Estimates		IPW Estimates	
	Point Est.	95% CI	Point Est.	95% CI	Point Est.	95% CI
$ATE(1,0)$	.054	$[-.010, .114]$	.054	$[-.009, .114]$	.052	$[-.011, .113]$
$NDE(1,0)$	.052	$[-.012, .112]$	.052	$[-.012, .111]$	.050	$[-.013, .110]$
$NIE(1,0)$	.002	$[-.002, .010]$	.002	$[-.002, .010]$	.002	$[-.002, .010]$
$CDE(1,0,4)$	.052	$[-.011, .111]$	.055	$[-.010, .115]$	.049	$[-.016, .112]$

Note: Estimates are expressed on probability scale (i.e., as risk differences). Bootstrap confidence intervals are based on $B = 2000$ replications. The linear model estimates are based on an additive linear model for $\mathbb{E}[M|C,D]$ and a linear model for $\mathbb{E}[Y|C,D,M]$ with a with $D \times M$ interaction. The simulation estimates are based on a normal linear model for $g(M|C,D)$, a logit model for $h(Y|C,D,M)$, and $J = 1000$ simulations. The inverse probability weighting (IPW) estimates are based on logit models for $f(D|C)$ and $s(D|C,M)$, and a normal linear model for $g(M|C,D)$. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/table_3-7`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/JOBSII`.

These effects can all be identified from the JOBSII study if the following conditions are satisfied: (i) there is no unobserved confounding of the exposure-outcome and exposure-mediator relationships; (ii) there is no unobserved or exposure-induced confounding of the mediator-outcome relationship; (iii) there is a positive probability of experiencing all levels of the exposure and mediator conditional on the baseline covariates; and (iv) the observed, potential, nested potential, and joint potential outcomes are consistent with one another. The first of these conditions is met by design in the JOBSII study because access to the job training workshop was randomly assigned. However, the second, third, and fourth conditions may not hold under the JOBSII study design, and they cannot be empirically verified. Nevertheless, we attempt to satisfy the second condition by adjusting for a set of baseline confounders. Additionally, structural violations of the third condition seem unlikely, and random departures from positivity in the sample data can be mitigated via modeling. The fourth condition, by contrast, is subject to greater doubt because any intervention that might be hypothetically deployed to manipulate self-efficacy is not clearly prescribed and could take multiple different forms.

We estimated the total, direct, and indirect effects defined previously using three different parametric approaches. First, we estimated these effects using a set of linear models for the mediator and outcome, as described in Section 3.5.1. The outcome model here includes an exposure-mediator interaction, but both of these models are otherwise additive in the predictors. Second, we estimated the effects of interest with the simulation approach described in Section 3.5.2. We implemented this approach using a normal linear model for the mediator, a logit model for the outcome (with an exposure-mediator interaction), and $J = 1000$ simulations. Finally, we computed a third set of effect estimates using inverse probability weights, as outlined in Section 3.5.3. To compute the weights, we fit logit models for the probability of assignment to the job training workshop, and we fit a normal linear model for the distribution of self-efficacy. These different approaches will yield consistent estimates if the models used to implement them are correctly specified and if the identification assumptions discussed previously are all satisfied in the JOBSII study.

The results of this analysis are reported in Table 3.7. Point estimates for the total effect suggest that job training increases the probability of reemployment by about 5 percentage points. Point estimates for the natural and controlled direct effects are comparable to those for the total effect, suggesting that workshop participation increases the chances of reemployment apart from any mechanism involving self-efficacy. By contrast, point estimates for the natural indirect effect are close to zero, which indicates that an individual's confidence in their ability to perform job search activities does not mediate the effect of job training. This pattern of results is highly consistent across the different estimation procedures we employed. However, all confidence intervals–computed using the nonparametric bootstrap with $B = 2000$ replications–are fairly wide and uniformly span zero.

Figure 3.10 presents a set of contour graphs summarizing the degree to which our estimates for the natural direct and indirect effects are sensitive to mediator-outcome confounding by an unobserved variable. Unobserved mediator-outcome confounding remains a concern in this analysis because only the exposure, and not the mediator, was randomly assigned in the JOBSII study. To assess the sensitivity of our estimates, we assumed the presence of a binary unobserved confounder U. For example, U might represent a participant's disability status at baseline, which could affect both their self-efficacy and their ability to secure new employment. We additionally assumed that the partial relationship of U with the outcome is invariant and that the difference in the probability of U associated with participation in the job training workshop is constant across levels of the mediator and confounders. Under these assumptions, Bias $\left(\widehat{NDE}\left(d, d^*\right)\right) = \delta_{UY|C,D,M} \times \delta_{DU|C,M}$ and Bias $\left(\widehat{NIE}\left(d, d^*\right)\right) = -\delta_{UY|C,D,M} \times \delta_{DU|C,M}$, where $\delta_{UY|C,D,M} = \mathbb{E}\left[Y|c, d, m, U = 1\right] - \mathbb{E}\left[Y|c, d, m, U = 0\right]$ and $\delta_{DU|C,M} = P\left(U = 1|c, d, m\right) - P\left(U = 1|c, d^*, m\right)$, exactly as in Section 3.7.3.

The contour graphs in Figure 3.10 plot bias-adjusted estimates across a range of plausible values for the sensitivity parameters $\delta_{UY|C,D,M}$ and $\delta_{DU|C,M}$. These estimates suggest that our findings are robust to mediator-outcome confounding by an unobserved variable. Specifically, across all values for the sensitivity parameters, bias-adjusted estimates for the natural indirect effect remain substantively small, while bias-adjusted estimates for the natural direct effect hover around 5 percentage points.

To summarize, our analysis of data from the JOBSII study provides relatively weak evidence that job training increases the chances of reemployment. Moreover, we also find little evidence that this effect, if present at all, is mediated by differences in job search self-efficacy. The data and code for replicating this analysis are available via hyperlinks in the footnotes to Table 3.7 and Figure 3.10.

3.9 Summary

In this chapter, we introduced methods for analyzing causal mediation in settings where there is a single mediator of interest and any potentially confounding variables are not exposure-induced. Using potential outcomes notation and the counterfactual framework, we showed how the average total effect of an exposure on an outcome can be decomposed into an indirect component operating through the mediator and direct component that is not transmitted through the mediator. We then explained how these effects can be nonparametrically identified from observable data under a set of assumptions, which require, among other things, that there is not any unobserved confounding of the exposure-outcome, exposure-mediator, or mediator-outcome relationships.

After defining our target estimands and explaining the conditions under which they can be identified, we turned our attention to estimation and inference. Specifically, we described several approaches to estimating direct and indirect effects using linear regression models and simple functions of their parameters, generalized linear models with simulation or imputation, and inverse probability weighting. The approach based on linear

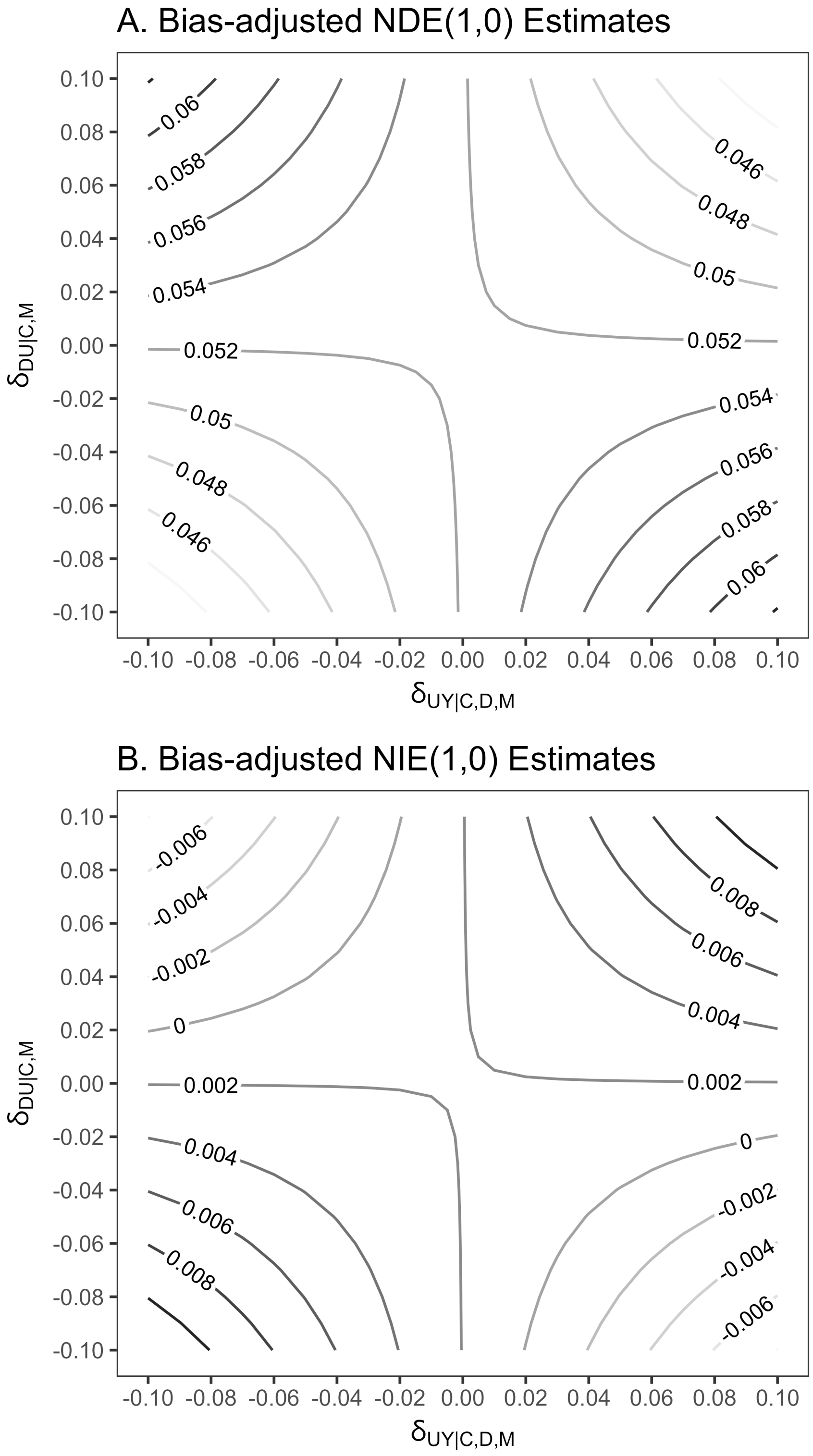


Figure 3.10: Estimates of Natural Direct and Indirect Effects from JOBSII Adjusted for Bias due to Unobserved Mediator-Outcome Confounding.

Note: Estimates are based on an additive linear model for $\mathbb{E}[M|C,D]$ and a linear model for $\mathbb{E}[Y|C,D,M]$ with a with $D \times M$ interaction. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/figure_3-10`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/JOBSII`.

regression models is best suited for applications with a continuous outcome and mediator, although there are situations where it may also perform reasonably well when these variables are binary, ordinal, or counts. The regression-based approach easily accommodates exposure-mediator interactions as well as interactions among the baseline confounders, exposure, and mediator, extending traditional approaches to mediation analysis in the social sciences. The simulation approach can be implemented with an even broader class of models for the mediator and outcome, allowing for many different types of nonlinearity, interactions, and response distributions. Inverse probability weighting requires models for the exposure rather than models for the mediator and outcome. It tends to perform best when the exposure is binary or otherwise takes on relatively small number of discrete values.

All these approaches to estimation require that there is not any unobserved confounding of the exposure-outcome, exposure-mediator, or mediator-outcome relationships, and also that there is not any exposure-induced confounders, whether observed or not. Otherwise, they will suffer from confounding bias. Additionally, these approaches require correct specification of the different models on which they are based. If these models are not correctly specified, the estimates they produce will suffer from misspecification bias, even if there is no unobserved or exposure-induced confounding. Researchers should therefore experiment with different model specifications and estimators, and they should consider implement a formal sensitivity analysis to assess the robustness of their inferences to hypothetical patterns of unobserved confounding.

This chapter elided several advanced topics in the interest of parsimony, including conditional effects and so-called moderated mediation, estimands involving ratios of potential outcomes, decompositions that further partition effects into components isolating different forms of interaction, and bias due to measurement error. We refer researchers interested in these topics to the specialized literature on conditional direct and indirect effects (VanderWeele 2015; Vansteelandt and VanderWeele 2012), alternative estimands based on relative risks or odds ratios (Samoilenko and Lefebvre 2019; Tchetgen Tchetgen 2014; VanderWeele and Vansteelandt 2010), three- and four-way decompositions of total effects (VanderWeele 2013, 2014), and adjustments for measurement error in the mediator (le Cessie et al. 2012; VanderWeele et al. 2012).

Chapter 4

Mediation Analysis with Exposure-induced Confounding

In the previous chapter, we explored the relationship between education and mental health, specifically examining whether college attendance reduces depression at midlife. We hypothesized that higher education would decrease the likelihood of unemployment, which, in turn, would reduce the risk of depression. In other words, we expected that unemployment would *mediate* the effect of college attendance. Using data from the 1979 National Longitudinal Survey of Youth (NLSY; Bureau of Labor Statistics 2019), we found that attending college does indeed reduce the chances of experiencing depression in adulthood, consistent with prior research on the link between education and mental health (Adams et al. 2003; Miech et al. 1999; Warren 2009; Yan and Williams 1999). However, we found little evidence that the negative effect of college attendance on depression is mediated by unemployment. This begs the question: if not through a mechanism involving unemployment, how does education reduce the risk of depression?

Another possibility is that attending college may improve mental health through its effects on income. Many prior studies have consistently documented a strong causal relationship between education and income, indicating that additional years of schooling lead to increased earnings (Angrist and Krueger 1991; Card 2001; Hout 2012). Formal education equips individuals with valuable skills, knowledge, and expertise, enhancing their productivity in the workplace and increasing the compensation employers are willing to offer them. Moreover, education also provides individuals with credentials that can give them an advantage in the job search process, enable access to well-paying occupations, and facilitate career advancement. As a result, educational attainment is a pivotal determinant of earning potential.

Higher incomes, in turn, may have a positive impact on mental health (Ridley et al. 2020; Shields-Zeeman and Smit 2022; Thomson et al. 2022). When individuals earn more money, they are better equipped to meet their basic consumption needs, adopt healthier lifestyles, and access quality healthcare, which reduces exposure to stressors and promotes psychological well-being. Higher incomes may also enhance social integration, boost self-esteem, and increase an individual's sense of control over their life, all of which are associated with improved mental health (Mirowsky and Ross 1990; Silverstone and Salsali 2003; Umberson and Karas Montez 2010). Thus, education may reduce the likelihood of depression by providing individuals with greater access to income.

How can we determine if income mediates the effect of post-secondary education on depression? Initially, we might consider using the methods discussed in Chapter 3, but with income as our focal mediator instead of unemployment. This approach seems straightforward, as we could easily substitute a measure of income for unemployment in our previous analyses and replicate them with this new mediator in place. However, as it turns out, this approach would be naive because it is needlessly susceptible to bias from exposure-induced confounding of the mediator-outcome relationship.

In particular, since education may affect the risk of unemployment, and unemployment directly influences income and potentially also mental health, it remains possible that unemployment is an exposure-induced confounder for the effect of income on depression. This type of confounding would introduce bias into the methods that we have discussed thus far for analyzing causal mediation. Although we found limited evidence of a causal chain linking college attendance with depression through the risk of unemployment, it would be unwise to dismiss this possibility entirely, especially if we can easily avoid it. So, how can we analyze whether income mediates the effect of college attendance on depression while accounting for potential exposure-induced confounders, such as unemployment status?

In this chapter, we introduce methods for analyzing whether a single mediator explains the effect of an exposure on an outcome in the presence of exposure-induced confounders. We begin by using causal graphs to present a series of mediation models that now incorporate multiple mediators, some of which may function as exposure-induced confounders. Next, we explain why natural direct and indirect effects cannot be nonparametrically identified when an exposure-induced variable confounds the effect of a focal mediator on the outcome. We then introduce a new set of estimands, known as interventional direct and indirect effects. Similar to natural direct and indirect effects, these interventional effects also capture how the influence of an exposure on an outcome is transmitted through a mediator of interest. However, unlike their natural counterparts, interventional direct and indirect effects can be nonparametrically identified and consistently estimated in the presence of exposure-induced confounders.

After outlining the conditions necessary for a nonparametric analysis of interventional effects, we shift our focus to parametric approaches to estimation. Specifically, we explain how linear models and the method of regression-with-residuals can be employed to consistently estimate interventional direct and indirect effects (Wodtke and Almirall 2017; Wodtke and Zhou 2020; Zhou and Wodtke 2019). In addition, we show how these effects can also be estimated using a broad class of generalized linear models (GLMs; Fox 2015; McCullagh 1989) and simulation methods. Lastly, we outline a third approach to estimation for interventional effects that involves re-weighting the data and then comparing differently weighted samples (VanderWeele 2009b; VanderWeele et al. 2014).

Throughout this chapter, we continue using data from the NLSY to illustrate key concepts and methods. We extend our analysis from Chapter 3 by examining whether income mediates the effect of college attendance on depression, now treating unemployment as a potential exposure-induced confounder.

We also present a second empirical illustration using data from Alesina et al. (2013). This study examined the influence of historical plow use on contemporary levels of female political participation in a cross-national analysis, and its findings highlighted the mediating role of a country's gross domestic product (GDP). To further explore whether GDP mediates the effect of historical plow use on women's involvement in politics, we reanalyze the data by accounting for a potentially important exposure-induced confounder–the degree to which a country's governance is authoritarian versus democratic.

Stata and R codes for implementing the analyses described in this chapter are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch4`. The footnotes accompanying each table and figure also contain hyperlinks to the specific scripts and data files used to generate their contents.

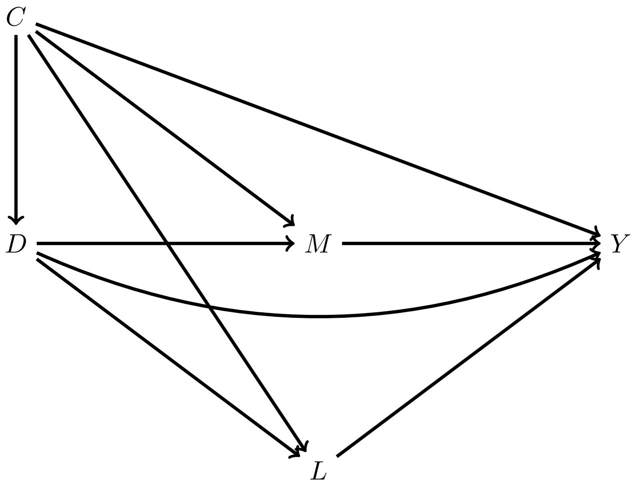


Figure 4.1: A Mediation Model with Baseline Confounding and Causally Unrelated Mediators.

Note: D denotes the exposure, M denotes a focal mediator of interest, L denotes a second mediator not of immediate interest, Y denotes the outcome, and C denotes a set of baseline confounders.

4.1 Graphical Models with Multiple Mediators

Figure 4.1 displays a graphical mediation model that incorporates multiple mediators. Specifically, it uses a directed acyclic graph (DAG; Elwert 2013; Pearl 2009) to describe causal relationships among a set of variables without detailing their functional form. Like the graphical models we discussed in Chapter 3, the model in Figure 4.1 encompasses an exposure D, an outcome Y, a mediator of interest M, and a collection of baseline confounders C. The arrows emanating from D to M, and then from M to Y, indicate that M partially mediates the impact of the exposure on the outcome. The main focus of this chapter, as before, is to measure the degree to which the effect of D on Y is mediated through M.

The graphical model in Figure 4.1 differs from those examined in the previous chapter, as it contains an additional variable, represented by L, that also mediates the effect of D on Y. The arrows from D to L, and then from L to Y, indicate that the exposure not only influences our mediator of interest M but also impacts another mediator L, which in turn affects the outcome. Consequently, the exposure now influences the outcome through three distinct mechanisms. It affects the outcome directly, as indicated by the $D \rightarrow Y$ path. It also affects the outcome indirectly through the mediator of interest M, as indicated by the $D \rightarrow M \rightarrow Y$ path. And it affects the outcome indirectly via the additional mediator L, as depicted by the $D \rightarrow L \rightarrow Y$ path.

An important characteristic of this model is that the two mediators are causally independent: M does not affect L, nor does L affect M. Because our mediator of interest M is not affected by the other mediator L, the model shown in Figure 4.1 does not involve exposure-induced confounding. Without any exposure-induced confounding, we can safely use the methods discussed in Chapter 3 to analyze whether and to what extent M mediates the effect of exposure D on the outcome Y. Furthermore, if we were interested in the mediating role of L instead, we could simply replace M with L as our focal mediator, and then we could apply the same methods from Chapter 3, now focusing on whether the effect of exposure D on the outcome Y is mediated by L rather than M.

With data stemming from a causal process like the model in Figure 4.1–where an exposure impacts an

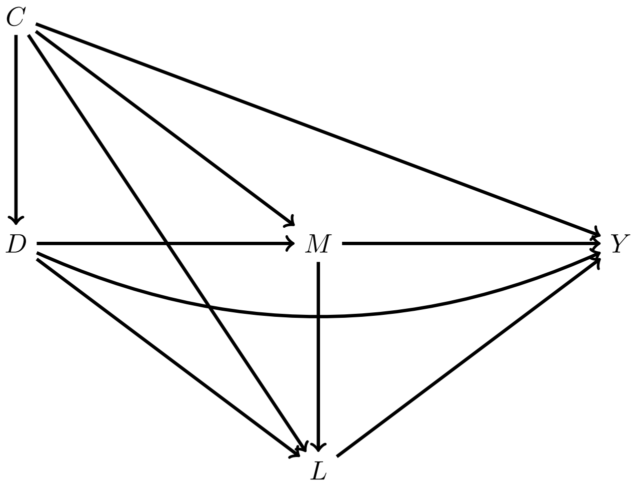


Figure 4.2: A Mediation Model with Baseline Confounding and Causally Ordered Mediators.

Note: D denotes the exposure, M denotes a focal mediator of interest, L denotes a second mediator not of immediate interest, Y denotes the outcome, and C denotes a set of baseline confounders.

outcome through multiple mediators that are causally unrelated–we can evaluate the contribution of each mediator to the total effect of exposure by focusing on one mediator at a time and applying the methods covered previously.

However, the model in Figure 4.1 is overly simplistic for many empirical applications in the social sciences. Because it is difficult to devise experiments that might ensure the causal independence of different mediators, applications encompassing multiple mediators will typically need to account for at least some form of causal interrelation among them. At the very least, outright dismissing the possibility that different mediators may affect one another is generally not advisable. When multiple mediators contribute to the effect of an exposure on an outcome, applying the methods from Chapter 3 to each mediator separately could lead to biased estimates and erroneous conclusions if any of these variables share a causal link.

Nevertheless, certain circumstances permit the continued use of methods covered previously in Chapter 3, even in applications with multiple mediators that are causally linked. Figure 4.2 depicts a graphical model nearly identical to the one in Figure 4.1 but with an additional arrow leading from our mediator of interest M to the other mediator L. Consequently, the mediators in this model share a causal relationship, where M affects L.

With this alteration, there are now three distinct pathways through which the exposure impacts the outcome indirectly. First, the exposure affects the outcome indirectly through M alone, as indicated by the $D \rightarrow M \rightarrow Y$ path. Second, it also affects the outcome indirectly through L alone, as indicated by the $D \rightarrow L \rightarrow Y$ path. Finally, the exposure affects the outcome indirectly through the combined action of M and L together, as represented by the $D \rightarrow M \rightarrow L \rightarrow Y$ path. In this path, our mediator of interest M precedes the other mediator L in a causal chain linking exposure D to the outcome Y, such that M is the first in a sequence of mediators that transmit the effect of exposure.

With data generated from a process resembling Figure 4.2, we can still safely employ methods from Chapter 3 to analyze how M mediates the effect of D on Y. Despite the presence of multiple, causally interrelated mediators, our focus is on the first link in the causal chain, which remains unaffected by exposure-induced confounding. As in Figure 4.1, the model in Figure 4.2 also lacks any variables that jointly affect

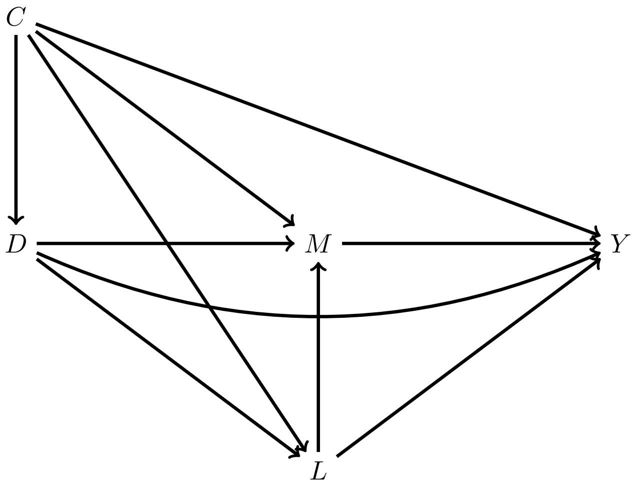


Figure 4.3: A Mediation Model with Baseline and Exposure-induced Confounding.

Note: D denotes the exposure, M denotes a focal mediator of interest, L denotes an exposure-induced confounder, Y denotes the outcome, and C denotes a set of baseline confounders.

both M and Y and that are also impacted by D. In this situation, we can continue using the methods covered previously to evaluate the mediating role of M.

However, if we were instead focused on the mediating role of L, we could not simply replace M with L as our focal mediator and apply the same methods from Chapter 3. This is because, in Figure 4.2, L is subject to exposure-induced confounding. Specifically, because M influences both L and Y and is also affected by D, it acts as an exposure-induced confounder for the relationship of L with Y. In this situation, naively applying the methods covered previously to assess the mediating role of L, rather than M, would lead to faulty inferences, as these methods presume the absence of exposure-induced confounding for the focal mediator.

Figure 4.3 displays a third graphical model with multiple mediators, but in this case, the causal order of the mediators, M and L, is reversed. In other words, this model is nearly identical to the model in Figure 4.2, but with one key difference: L affects M, rather than the other way around. The two mediators in this model therefore maintain a causal relationship, only now because our mediator of interest M is affected by the other mediator L. As a result, the link between M and the outcome Y suffers from exposure-induced confounding due to L. With data generated from a model like this one, we could safely use the methods covered in Chapter 3 to analyze how L mediates the effect of D on Y, but we could not use these methods to analyze the mediating role of M, contaminated as it is by exposure-induced confounding.

To summarize, the methods from Chapter 3 can be reliably applied to evaluate the explanatory role of a single mediator, even when multiple mediators transmit the effect of exposure, as long as these mediators are causally unrelated or, short of that, the mediator of interest precedes the others in causal order. Otherwise, the focal mediator will suffer from exposure-induced confounding, rendering these methods generally unsuitable.

In this chapter, we introduce methods for analyzing mediation with data generated from a process that resembles Figure 4.3. Specifically, with data of this sort, we focus on evaluating the role of a single mediator of interest M in transmitting the effect of an exposure D on an outcome Y, when exposure-induced confounders like L are present. We defer consideration of methods for assessing the explanatory role of multiple mediators simultaneously until Chapter 5, where we examine how to analyze mediation operating through a set of different variables acting together or in sequence. For now, our focus is limited to methods for isolating the explanatory role of a single mediator when its effect on the outcome is confounded by other mediators.

In our empirical illustration using the NLSY, the exposure D denotes college attendance by age 22, the outcome Y denotes standardized scores on the Center for Epidemiological Studies-Depression Scale (CES-D; Radloff 1977) at age 40, and the baseline confounders C include measures of race, gender, parental education, and so on, exactly as in Chapter 3. The mediator of interest M now represents a sample member's household income measured between age 35 to 39, whereas their prior experience with unemployment, denoted in this chapter by L, is treated as an exposure-induced confounder.

We assume that these data stem from a causal process resembling Figure 4.3, where college attendance may affect depression through several causal pathways. It may influence depression directly or through unobserved mechanisms, as indicated by the $D \to Y$ path. College attendance may also influence depression indirectly through its impact on unemployment and income. For example, attending college may affect an individual's likelihood of unemployment. In turn, the experience of unemployment could directly influence depression, as indicated by the $D \to L \to Y$ path, or it could indirectly affect depression through its impact on income, as shown by the $D \to L \to M \to Y$ path. Furthermore, college attendance might also influence depression indirectly by affecting income through other mechanisms beyond the risk of unemployment, as indicated by the $D \to M \to Y$ path.

Our objective is to isolate the mediating role of income from the potentially confounding influence of both baseline characteristics and unemployment. The confounding influence of baseline characteristics is graphically depicted by the $D \leftarrow C \to Y$, $D \leftarrow C \to M$, $M \leftarrow C \to Y$ and $M \leftarrow L \leftarrow C \to Y$ paths, whereas the confounding influence of unemployment is represented by the $M \leftarrow L \to Y$ path. Any potential confounding by unemployment is considered exposure-induced, due to the presence of a causal path from D to L.

As discussed previously, it is important to remember that we cannot definitively ascertain whether the data truly stem from a causal process resembling Figure 4.3. Rather, this is an assumption. If the actual causal process differs from that assumed, the methods in this chapter could lead to inaccurate conclusions regarding the explanatory role of a focal mediator. We formally outline the assumptions required of these methods below, and revisit them throughout the chapter, highlighting instances where they may not hold true and the resulting implications.

4.2 Limitations of the Natural Effects Decomposition

In the previous chapter, we analyzed causal mediation by decomposing an average total effect of the exposure D on the outcome Y into two distinct components: a natural indirect effect operating through a mediator of interest M, and a natural direct effect operating through other mechanisms. The average total effect is formally defined as $ATE\left(d, d^*\right) = \mathbb{E}\left[Y\left(d\right) - Y\left(d^*\right)\right]$. It represents the expected difference in the outcome arising from exposure to d rather than d^*. This effect can be separated into direct and indirect components as follows: $ATE\left(d, d^*\right) = NDE\left(d, d^*\right) + NIE\left(d, d^*\right)$, which we refer to here as the *natural effects decomposition.*

In this decomposition, the natural indirect effect is defined as $NIE\left(d, d^*\right) = \mathbb{E}\left[Y\left(d, M\left(d\right)\right) - Y\left(d, M\left(d^*\right)\right)\right]$. It captures a component of the total effect of exposure on the outcome that is transmitted through the mediator of interest. The natural direct effect, on the other hand, is defined as $NDE\left(d, d^*\right) = \mathbb{E}\left[Y\left(d, M\left(d^*\right)\right) - Y\left(d^*, M\left(d^*\right)\right)\right]$. It captures a component of the total effect that is not transmitted through the mediator of interest.

In these expressions, recall that $Y(d)$ and $M(d)$ denote potential values of the outcome and mediator, respectively, if an individual were exposed to d. They represent what the outcome and mediator would have been if an individual experienced this particular level of the exposure, possibly contrary to fact. The term $Y(d, M(d))$ denotes a nested potential outcome. It represents the value of the outcome that would arise if an individual were exposed to d and, by extension, to the level of the mediator that they would naturally experience under exposure d. Similarly, $Y(d, M(d^*))$ denotes a cross-world potential outcome. This term represents the value of the outcome that would arise if an individual were exposed to d but then experienced the level of the mediator that would have naturally arisen for them under the alternative exposure d^*.

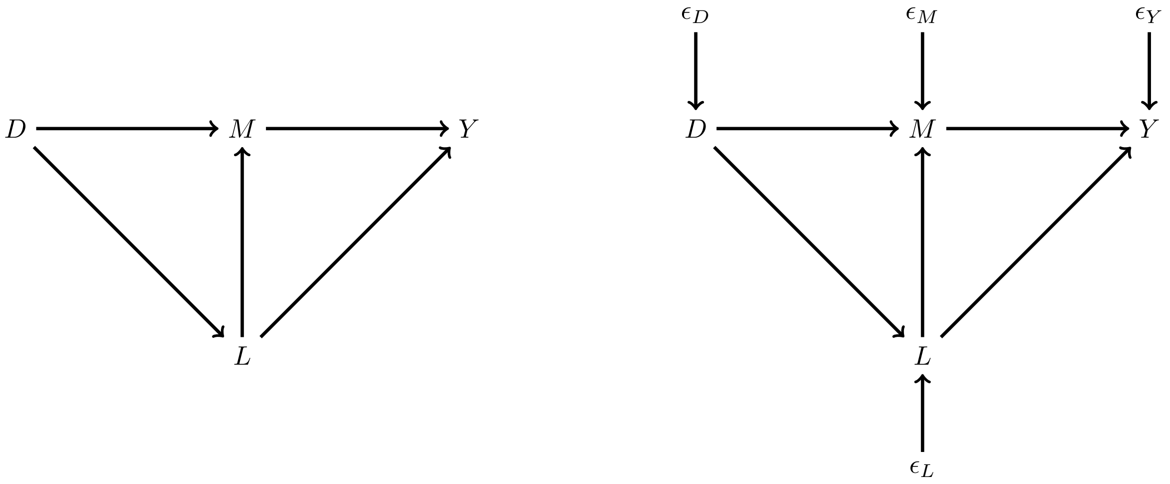
The nested and cross-world potential outcomes are both a special type of joint potential outcome, which represents the outcome value that would arise for an individual under particular levels of the exposure and mediator taken together. A joint potential outcome is denoted by $Y(d, m)$. It captures what the outcome would have been if an individual experienced level d of the exposure and level m of the mediator.

Although natural direct and indirect effects offer a convenient way to differentiate the impact of an exposure operating through a mediator of interest versus alternative mechanisms, their nonparametric identification hinges on meeting a set of stringent conditions. Specifically, to nonparametrically identify the $NDE(d, d^*)$ and $NIE(d, d^*)$, we must satisfy the following conditional independence assumptions, as outlined in the previous chapter. First, assumption (c.i) requires that the exposure must be independent of the joint potential outcomes, given the baseline confounders. This assumption is formally expressed as $Y(d, m) \perp D|C$. Second, assumption (c.ii) requires that the mediator must be independent of the joint potential outcomes, given the baseline confounders and exposure to d. This assumption is denoted formally as $Y(d, m) \perp M|C, D = d$. Third, assumption (c.iii) requires that the exposure must be independent of the potential values for the mediator, given the baseline confounders. We represent this assumption formally as $M(d) \perp D|C$. Lastly, assumption (c.iv) requires that the joint potential outcomes under exposure to d must be independent of the potential values for the mediator under exposure to d^*, given the baseline confounders. We represent this assumption formally as $Y(d, m) \perp M(d^*)|C$. Known as a cross-world independence assumption, this condition is especially restrictive because it stipulates that the joint potential outcomes must be unrelated to potential values of the mediator across different counterfactual worlds.

A significant drawback of the natural effects decomposition is that its components cannot be nonparametrically identified in the presence of exposure-induced confounders, whether these variables are observed or not. This is because the presence of a mediator-outcome confounder influenced by exposure will inevitably violate the assumption of cross-world independence (c.iv). This violation precludes the identifiability of natural direct and indirect effects, unless other and even stricter assumptions are invoked. Since exposure-induced confounders are common in analyses of causal mediation, the applicability of the natural effects decomposition is limited by its reliance on the cross-world independence assumption, which renders this approach unsuitable in many contexts.

To appreciate why the presence of an exposure-induced confounder prevents identification for natural direct and indirect effects, consider the DAG in Panel A of Figure 4.4. This figure describes a simple mediation model with exposure-induced confounding.

In this model, the exposure D is not confounded, and it directly influences the mediator of interest M. The mediator M, in turn, directly influences the outcome Y. However, the effect of M on Y is confounded by another variable L, which is influenced by the exposure D. Thus, L is an exposure-induced confounder of the mediator-outcome relationship.



A. Conventional DAG

B. DAG with Random Disturbances

Figure 4.4: A Simple Mediation Model with Exposure-induced Confounding.

Note: D denotes the exposure, M denotes a focal mediator of interest, L denotes an exposure-induced confounder, and Y denotes the outcome. The $\{\epsilon_D, \epsilon_L, \epsilon_M, \epsilon_Y\}$ terms are a set of random disturbances suppressed from Panel A for visual simplicity but depicted explicitly in Panel B for illustration. The two graphs are functionally equivalent.

As outlined in Chapter 2, a DAG can also be expressed as a system of nonparametric structural equations, where each variable is determined by a generic function of its causal antecedents and a random disturbance. Specifically, the graph in Figure 4.4 can be translated into the following system of equations:

$$\begin{aligned} D &\coloneqq f_D\left(\epsilon_D\right) \\ L &\coloneqq f_L\left(D, \epsilon_L\right) \\ M &\coloneqq f_M\left(D, L, \epsilon_M\right) \\ Y &\coloneqq f_Y\left(L, M, \epsilon_Y\right). \end{aligned} \tag{4.1}$$

In these equations, $\{\epsilon_D, \epsilon_L, \epsilon_M, \epsilon_Y\}$ denotes a set of random disturbances, and $\{f_D, f_L, f_M, f_Y\}$ represents a set of functions that do not encode any restrictions on the form of the causal relationships among variables. The $\coloneqq$ symbol is an assignment operator, which signals the direction of causal influence. Note that while random disturbances are always implicitly present in a DAG, they are typically suppressed from these diagrams for visual parsimony. However, if desired, they can be visually integrated into the graph for a more complete representation, as shown in Panel B of Figure 4.4.

In Equation 4.1, $D \coloneqq f_D\left(\epsilon_D\right)$ signifies that the exposure D is determined by an unrestricted function $f_D\left(\cdot\right)$ of a random disturbance ϵ_D. For example, ϵ_D might be the outcome of a coin flip in an experiment where exposure is randomly assigned. Similarly, $L \coloneqq f_L\left(D, \epsilon_L\right)$ indicates that the confounder L is determined by some function $f_L\left(\cdot\right)$ of the exposure D and a random disturbance ϵ_L. Next, $M \coloneqq f_M\left(D, L, \epsilon_M\right)$ indicates that the mediator M is determined by a function $f_M\left(\cdot\right)$ of the exposure D, confounder L, and a random disturbance ϵ_M. Finally, $Y \coloneqq f_Y\left(L, M, \epsilon_Y\right)$ denotes that the outcome is determined by a function $f_Y\left(\cdot\right)$ of the confounder L, mediator M, and a random disturbance ϵ_Y. Taken together, these expressions encapsulate the same information that is visually depicted in the DAG, translating a graphical representation with nodes and arrows into mathematical notation.

Given the system of causal relationships in Equation 4.1, we can construct potential values for the confounder L, mediator of interest M, and outcome Y as follows:

$$\begin{aligned} L\left(d\right) &\coloneqq f_L\left(d, \epsilon_L\right) \\ L\left(d^*\right) &\coloneqq f_L\left(d^*, \epsilon_L\right) \\ M\left(d^*\right) &\coloneqq f_M\left(d^*, L\left(d^*\right), \epsilon_M\right) = f_M\left(d^*, f_L\left(d^*, \epsilon_L\right), \epsilon_M\right) \\ Y\left(d, m\right) &\coloneqq f_Y\left(L\left(d\right), m, \epsilon_Y\right) = f_Y\left(f_L\left(d, \epsilon_L\right), m, \epsilon_Y\right). \end{aligned} \tag{4.2}$$

These expressions illustrate why exposure-induced confounding precludes nonparametric identification for natural direct and indirect effects. Specifically, Equation 4.2 shows that the assumption of cross-world independence between $Y\left(d, m\right)$ and $M\left(d^*\right)$ is violated, as both these variables share a common influence, ϵ_L, the disturbance term for the exposure-induced confounder. In other words, $Y\left(d, m\right)$ is related to $M\left(d^*\right)$ because both variables are partly determined by the random component of L, and this relation breaches a critical condition required to nonparametrically identify the natural effects decomposition.

Moreover, $Y\left(d, m\right)$ is not even independent of $M\left(d^*\right)$ within levels of the exposure-induced confounder. Whether or not we condition on the observed value of the exposure-induced confounder L, its potential values, $L\left(d^*\right)$ and $L\left(d\right)$, remain intertwined due to their shared disturbance term ϵ_L. And because the potential values of the mediator and outcome, $M\left(d^*\right)$ and $Y\left(d, m\right)$, depend on $L\left(d^*\right)$ and $L\left(d\right)$ in turn, they too remain interdependent, regardless of whether we condition on L (Andrews and Didelez 2021). Thus, natural direct and indirect effects cannot be nonparametrically identified in the presence of exposure-induced confounders, even when these variables are observed and available for statistical control.

Although nonparametric identification of natural direct and indirect effects is a nonstarter in the presence of exposure-induced confounders, parametric approaches to identification remain viable. For example, in addition to assumptions (c.i) to (c.iii), we could also assume that $f_L\left(D, \epsilon_L\right)$, $f_M\left(D, L, \epsilon_M\right)$, and $f_Y\left(L, M, \epsilon_Y\right)$ are all linear and additive. Should these additional assumptions hold true, natural direct and indirect effects are identified by relatively simple functions of the coefficients in these linear and additive equations, despite the presence of an exposure-induced confounder (Andrews and Didelez 2021). This is the approach to identification employed in applications of linear path models–often implicitly–whenever multiple, causally related mediators are analyzed together (Alwin and Hauser 1975; Duncan 1966; Hayes 2017).

Although certain parametric assumptions, such as linearity and additivity, can facilitate identification for natural direct and indirect effects (Andrews and Didelez 2021; De Stavola et al. 2015; Robins and Greenland 1992), these assumptions often impose severe restrictions on the probability distribution of the data that cannot be substantiated by theory or prior scientific knowledge. This is especially concerning in the social sciences, where the empirical implications of our theories are comparatively vague and it is typically difficult to imagine causal relations among humans adhering to some simplistic functional form. Thus, relying on stringent parametric assumptions to identify the natural effects decomposition is not generally advisable, at least not without a compelling substantive rationale. While this approach may appear attractive due to its simplicity, our inferences about causal mediation will be mistaken should any of these parametric assumptions fail to hold.

Instead of employing arbitrary and often dubious parametric assumptions in an attempt to salvage the natural effects decomposition, a more defensible strategy might be to execute a mediation analysis that targets an alternate set of estimands. These alternative estimands, known as interventional direct and indirect effects, capture the operation of a causal chain transmitting the effect of exposure on an outcome through a specific mediator of interest, much like the natural effects decomposition. However, unlike natural direct and indirect effects, interventional effects can be nonparametrically identified and flexibly estimated under weaker assumptions that allow for the existence of exposure-induced confounders. This is the approach on which we focus throughout the rest of the chapter.

4.3 Interventional Direct and Indirect Effects

In this section, we define a set of causal estimands that can be used to evaluate mediation when the data arise from a process resembling Figure 4.3. Our focus is on evaluating causal mediation with a single mediator of interest, a set of baseline confounders, and one or more exposure-induced confounders. We concentrate specifically on interventional direct and indirect effects, as proposed by Geneletti (2007), Didelez et al. (2012), and VanderWeele et al. (2014). These effects are similar to their natural counterparts except for one significant difference: instead of setting the mediator to the level it would have naturally assumed for each individual under a particular exposure, interventional direct and indirect effects set the mediator at a value randomly drawn from its distribution under a given exposure. Defining these effects in terms of random draws from a mediator distribution breaks the association between potential values of the mediator and outcome due to the shared influence of an exposure-induced confounder, and it obviates the need for a cross-world independence assumption.

As in prior chapters, we define our target estimands using potential outcomes notation (Holland 1986; Rubin 1974). Let $Y(d)$ and $M(d)$ again denote the values of the outcome and mediator, respectively, that would have been observed for an individual if they had previously been exposed to d, possibly contrary to fact. Similarly, let $Y(d,m)$ denote the joint potential outcome for an individual if they had experienced level d of the exposure and level m of the mediator together.

Next, let $\mathcal{M}(d|C)$ denote a value of the mediator randomly selected from its distribution under exposure d, given the baseline confounders C. To elaborate, imagine a scenario where we assigned every individual in a target population to exposure d and then subsequently measured their mediator. If we grouped these individuals into subpopulations with identical levels on the baseline confounders, and then pooled all their mediator values together, these sets of values would compose the distribution of the mediator under exposure d, given the baseline confounders C. We use $\mathcal{M}(d|C)$ to denote a randomly selected value from this distribution. The script "M" differentiates between an individual's potential value of the mediator, denoted by $M(d)$, and a random draw from the conditional distribution of these potential values (Nguyen et al. 2022).

Finally, let $Y(d,\mathcal{M}(d|C))$ denote the value of the outcome that would have occurred for an individual if they had been exposed to d and then experienced a level of the mediator randomly selected from its distribution under exposure d. We refer to this type of potential outcome as a *randomized potential outcome* because it involves setting the mediator to a value randomly drawn from a distribution.

Using this notation, the *interventional direct effect* of an exposure D on an outcome Y can be formally defined as follows:

$$IDE(d,d^*) = \mathbb{E}\left[Y(d,\mathcal{M}(d^*|C)) - Y(d^*,\mathcal{M}(d^*|C))\right]. \tag{4.3}$$

This effect represents the average difference in the outcome if individuals had been exposed to d rather than d^*, and had then experienced a level of the mediator that was randomly selected from its distribution under exposure d. The $IDE(d,d^*)$ captures an effect of the exposure on the outcome that bypasses the mediator of interest. This is achieved by comparing outcomes under different levels of the exposure, d versus d^*, while holding the mediator constant at a value chosen randomly from its distribution under the reference level of the exposure, $\mathcal{M}(d^*|C)$. Thus, the interventional direct effect is also sometimes referred to as a "randomized intervention analogue of the natural direct effect" (e.g., VanderWeele et al. 2014; VanderWeele 2015; Wodtke and Zhou 2020).

Similarly, the *interventional indirect effect* of an exposure D on an outcome Y can be formally defined as follows:

$$IIE(d, d^*) = \mathbb{E}\left[Y(d, \mathcal{M}(d|C)) - Y(d, \mathcal{M}(d^*|C))\right]. \tag{4.4}$$

This effect represents the average difference in the outcome if individuals had been exposed to d and had then experienced a level of the mediator randomly selected from its distribution under exposure d as opposed to its distribution under exposure d^*. The $IIE(d, d^*)$ isolates an effect of the exposure on the outcome that operates through the mediator of interest. This is achieved by fixing the exposure at d and then comparing outcomes under different levels of the mediator, $\mathcal{M}(d|C)$ versus $\mathcal{M}(d^*|C)$, which are values chosen at random from two distinct distributions. The interventional indirect effect is also sometimes referred to as a "randomized intervention analogue of the natural indirect effect."

Interventional direct and indirect effects sum to equal a randomized intervention analogue of the average total effect. We refer to this analogue as the *overall effect* of an exposure D on an outcome Y (Nguyen et al. 2022; VanderWeele et al. 2014), which can be formally expressed as follows:

$$\begin{aligned} OE(d, d^*) &= IDE(d, d^*) + IIE(d, d^*) \\ &= \mathbb{E}\left[Y(d, \mathcal{M}(d^*|C)) - Y(d^*, \mathcal{M}(d^*|C))\right] + \mathbb{E}\left[Y(d, \mathcal{M}(d|C)) - Y(d, \mathcal{M}(d^*|C))\right] \\ &= \mathbb{E}\left[Y(d, \mathcal{M}(d|C)) - Y(d^*, \mathcal{M}(d^*|C))\right]. \end{aligned} \tag{4.5}$$

The $OE(d, d^*)$ mirrors an average total effect except that it is defined by contrasting different levels of the exposure and different levels of the mediator randomly selected from its distribution under each of the contrasted exposures. Specifically, it represents the average difference in the outcome if individuals had been exposed to d rather than d^* and if they had experienced a level of the mediator randomly selected from its distribution under exposure d as opposed to its distribution under exposure d^*.

Thus, as with the natural effects decomposition, the degree to which an exposure D affects an outcome Y through a mediator of interest M can also be evaluated with interventional direct and indirect effects. The $IDE(d, d^*)$ captures an effect of the exposure on the outcome that does not operate through its impact on the distribution of a focal mediator. In contrast, the $IIE(d, d^*)$ captures an effect on the outcome that arises from a shift in the distribution of the mediator caused by changes in the exposure. And the $OE(d, d^*)$, by extension, captures the joint effect of a change in the exposure and a corresponding shift in the distribution of the mediator arising from this change in exposure.

Despite their similarities, interventional effects differ from their natural counterparts in subtle but important ways. With natural direct and indirect effects, the mediator assumes whatever value it would have naturally been for each individual under a particular level of the exposure. With interventional direct and indirect effects, however, an individual's mediator is assigned a random value from its distribution under a specific exposure level. Because of this distinction, natural and interventional effects will only coincide under special circumstances.

One such circumstance is when there is no exposure-mediator interaction. When there is no interaction effect between the exposure and mediator on the outcome, the impact of different exposures is invariant regardless of whether the mediator for each individual is set to its potential value $M(d)$, a random draw from the distribution of these potential values $\mathcal{M}(d|C)$, or any other value m. In this situation, interventional direct and indirect effects are equal to their natural counterparts, and by extension, the overall effect coincides with the average total effect.

Furthermore, interventional effects will also align with their natural counterparts when there is no exposure-induced confounding. In applications with a single mediator of interest and baseline confounding only, this implies that natural direct and indirect effects can be interpreted as interventional effects, and vice versa. It also follows that the overall effect is equivalent to the average total effect in such applications. Thus, all our results based on the natural effects decomposition in Chapter 3 could be reinterpreted as interventional effects, provided our assumption of no exposure-induced confounding held true in those analyses.

While interventional effects conveniently align with their natural counterparts under certain conditions, they are fundamentally distinct estimands nonetheless. Natural effects are defined by cross-world potential outcomes that involve setting the exposure and mediator at values that can never be observed together. As a result, these effects cannot be mapped to feasible experiments, not even hypothetically. In contrast, interventional effects avoid these limitations, as they are not defined in terms of cross-world potential outcomes that are impossible to observe. Rather, they are defined in terms of randomized potential outcomes that are measurable, at least in theory. Interventional effects can therefore be identified under less stringent conditions than their natural counterparts, and they can be related to hypothetical experiments that could be performed in the real world. We elaborate on these points in Section 4.4 below.

Throughout this chapter, we also maintain our focus on controlled direct effects, alongside the $IDE(d, d^*)$, $IIE(d, d^*)$, and $OE(d, d^*)$. As discussed in Chapter 3, the controlled direct effect is denoted by $CDE(d, d^*, m) = \mathbb{E}[Y(d, m) - Y(d^*, m)]$. It represents the average difference in the outcome if individuals were exposed to d rather d^* but then experienced the same level of the mediator m. Like interventional direct and indirect effects, controlled direct effects can also be nonparametrically identified in the presence of exposure-induced confounders.

Interventional and controlled direct effects share other similarities. In fact, the interventional direct effect is just a weighted average of many different controlled direct effects. To appreciate this, note that the $IDE(d, d^*)$ can also be expressed as follows:

$$\begin{aligned} IDE(d, d^*) &= \mathbb{E}[Y(d, \mathcal{M}(d^*|C)) - Y(d^*, \mathcal{M}(d^*|C))] \\ &= \sum_{m,c} \mathbb{E}[Y(d, m) - Y(d^*, m)|c]\, P(M(d^*) = m|c)\, P(c) \\ &= \sum_{m,c} CDE(d, d^*, m|c)\, P(M(d^*) = m|c)\, P(c), \end{aligned} \tag{4.6}$$

where $CDE(d, d^*, m|c)$ denotes a controlled direct effect within the subpopulation of individuals for whom $C = c$. The final expression in Equation 4.6 is a weighted average of these controlled direct effects taken over the interventional distribution of the mediator under exposure to d^*.

In this section, we have provided definitions for interventional effects that are not reliant on a specific model or level of measurement for the exposure, mediator, and outcome. These definitions apply universally, regardless of the underlying process that generated the data.

To illustrate them more concretely, however, consider our empirical example based on the NLSY. With these data, the interventional direct effect of college attendance on depression can be expressed as $IDE(1, 0) = \mathbb{E}[Y(1, \mathcal{M}(0|C)) - Y(0, \mathcal{M}(0|C))]$. This effect represents the average difference in CES-D scores at age 40 if sample members had, versus had not, attended college by age 22 and if they then received an income between age 35 to 39 randomly selected from the distribution that would have arisen had they not attended college. The $IDE\left(1,0\right)$ captures an effect of college attendance on depression that is not due to differences in the distribution of income caused by changes in prior educational attainment.

Similarly, the interventional indirect effect can be expressed as $IIE\left(1,0\right) = \mathbb{E}\left[Y\left(1,\mathcal{M}\left(1|C\right)\right) - Y\left(1,\mathcal{M}\left(0|C\right)\right)\right]$ in the NLSY. This effect represents the average difference in CES-D scores if individuals had attended college and then received an income randomly selected from the distribution under this exposure instead of from the distribution that would have arisen had they not attended college. The $IIE\left(1,0\right)$ captures an effect of college attendance on depression that operates through differences in the distribution of income resulting specifically from changes in prior educational attainment.

The overall effect, obtained by summing the interventional direct and indirect effects, can be expressed as $OE\left(1,0\right) = \mathbb{E}\left[Y\left(1,\mathcal{M}\left(1|C\right)\right) - Y\left(0,\mathcal{M}\left(0|C\right)\right)\right]$. This effect represents the average difference in CES-D scores attributable to attending college and receiving an income randomly selected from the distribution under this higher level of education, compared to not attending college and receiving an income drawn from the distribution under this lower level of education. It captures the combined impact of college attendance and the resulting shift in the income distribution that would occur if everyone achieved at least some post-secondary education.

Finally, the controlled direct effect in our empirical example can be expressed as $CDE\left(1,0,m\right) = \mathbb{E}\left[Y\left(1,m\right) - Y\left(0,m\right)\right]$. With a continuous mediator like income, there are many different controlled direct effects, each corresponding to a specific value of m. We focus on an effect that controls incomes at \$50,000, which is approximately equal to the sample mean in the NLSY. This particular effect, formally denoted by $CDE\left(1,0,50\text{K}\right) = \mathbb{E}\left[Y\left(1,50\text{K}\right) - Y\left(0,50\text{K}\right)\right]$, represents the average difference in CES-D scores at age 40 if individuals had, versus had not, attended college by age 22 and had earned an annual income of \$50,000 between age 35 to 39. It captures how attending college would impact depression, even if everyone earned a "middle-class" income regardless of their educational background.

4.4 Nonparametric Identification

As discussed previously, nonparametric identification involves a process by which causal estimands defined in terms of counterfactuals are linked with observable population data. This link is established through a series of assumptions that do not impose any functional form restrictions on the probability distribution of these data (Hernan and Robins 2020; Lundberg et al. 2021). In this section, we explain how each of the estimands outlined previously can be nonparametrically identified.

4.4.1 Nonparametric Identification of Controlled Direct Effects

In the presence of exposure-induced confounders, nonparametric identification of controlled direct effects hinges on assumptions that align with those needed to achieve identification when confounding is due to baseline factors only. However, these assumptions are modified slightly to accommodate the additional confounding variables and their sensitivity to prior exposures. Specifically, the $CDE\left(d,d^{*},m\right)$ can be nonparametrically identified under the following assumptions when exposure-induced confounders are present: (d.i) conditional independence of the exposure with respect to the joint potential outcomes, (d.ii) conditional independence of the mediator with respect to the joint potential outcomes, (d.iii) sequential positivity of the

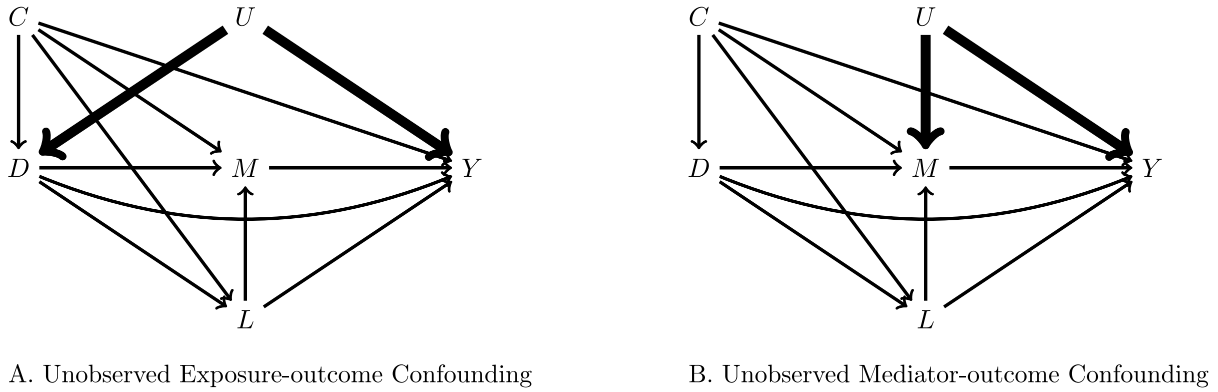

C
U
D
M
Y
L
A. Unobserved Exposure-outcome Confounding
C
U
D
M
Y
L
B. Unobserved Mediator-outcome Confounding

Figure 4.5: Graphical Illustration of Unobserved Exposure-outcome and Mediator-outcome Confounding.

Note: D denotes the exposure, M denotes a focal mediator of interest, L denotes an exposure-induced confounder, Y denotes the outcome, C denotes a set of baseline confounders, and U denotes a set of unobserved confounders.

exposure and mediator, and (d.iv) consistency of the observed and potential outcomes (Acharya et al. 2016; VanderWeele 2009b; Zhou and Wodtke 2019).

Assumption (d.i). The first of these assumptions requires that the exposure is statistically independent of the joint potential outcomes, conditional on the baseline confounders. This assumption can be expressed formally as $Y(d,m) \perp D|C$. It is identical to assumption (b.i) from Chapter 3, where we focused on analyzing controlled direct effects without exposure-induced confounding. In substantive terms, this assumption implies that there should be no unobserved confounding of the relationship between the exposure D and outcome Y, after we have accounted for the baseline confounders C.

Assumption (d.ii). Nonparametric identification of the $CDE(d,d^*,m)$ also requires that the mediator is statistically independent of the joint potential outcomes, conditional on the baseline confounders, the exposure, and the exposure-induced confounders. Formally, this assumption can be expressed as follows:

$$Y(d,m) \perp M|C,D,L. \tag{4.7}$$

Substantively, it implies that there must not be any unobserved confounding of the relationship between the mediator of interest M and the outcome Y, after we have accounted for the baseline confounders C, the exposure D, and the exposure-induced confounders L. Equation 4.7 resembles assumption (b.ii) from Chapter 3, but it additionally incorporates a set of mediator-outcome confounders L, which may be influenced by the exposure D.

Assumptions (d.i) and (d.ii) would both hold true in data generated from a process resembling Figure 4.3, where there are no unobserved variables that jointly affect D and Y or M and Y. However, if the data were generated from a process resembling Panel A of Figure 4.5, where an unobserved variable U affects both D and Y, assumption (d.i) would be violated. Similarly, assumption (d.ii) would be violated if the data were generated from a process resembling Panel B of Figure 4.5, where an unobserved variable U jointly affects M and Y.

In sum, nonparametric identification of the $CDE(d,d^*,m)$ requires that there is no exposure-outcome or mediator-outcome confounding by unobserved variables. Experimental studies that jointly or sequentially randomize the exposure and mediator would satisfy these conditions by design, since randomization ensures that the manipulated variables are independent of all other causes of the outcome, whether observed or not.

Assumption (d.iii). Nonparametric identification of the $CDE\,(d, d^{*}, m)$ also requires a positive probability for all values for the exposure, conditional on the baseline confounders. Additionally, it requires a positive probability for all values for the mediator, conditional on the baseline confounders, the exposure, and the exposure-induced confounders. Formally, this assumption can be expressed as follows:

$$P\,(d|c) > 0 \text{ and } P\,(m|c, d, l) > 0. \tag{4.8}$$

Substantively, it implies that there must be some chance for individuals to experience all levels of the exposure within every subpopulation defined by the baseline confounders. It also implies that there must be a chance for individuals to experience all levels of the mediator within every subpopulation defined not only by the baseline confounders but also by the exposure and exposure-induced confounders. These conditions resemble assumption (b.iii) from Chapter 3, except they require that positivity holds sequentially for the exposure and mediator, conditional on each of their antecedents.

Assumption (d.iv). Lastly, nonparametric identification of the $CDE\,(d, d^{*}, m)$ requires that the observed outcomes are consistent with the joint potential outcomes under the actual levels of the exposure and mediator that individuals experienced. This assumption can be formally expressed as $Y = Y\,(D, M)$,which is identical to assumption (b.iv) from Chapter 3. Substantively, it implies that there must not be multiple versions of the exposure or mediator with heterogeneous effects on the outcome or any interference between individuals.

Nonparametric identification formula. When assumptions (d.i) to (d.iv) are satisfied, controlled direct effects can be equated with observable data as follows:

$$CDE\,(d, d^{*}, m) = \sum_{l,c} \left(\mathbb{E}\,[Y|c, d, l, m]\,P\,(l|c, d) - \mathbb{E}\,[Y|c, d^{*}, l, m]\,P\,(l|c, d^{*})\right) P\,(c)\,. \tag{4.9}$$

We refer to this expression as the nonparametric identification formula for the $CDE\,(d, d^{*}, m)$ in the presence of exposure-induced confounders L. In Appendix G, we provide a step-by-step derivation of this identification formula, beginning with a definition of the $CDE\,(d, d^{*}, m)$ in terms of counterfactuals and concluding with the function of observable data in Equation 4.9.

The nonparametric identification formula involves comparing two different quantities. The first can be expressed as follows:

$$\sum_{l,c} \mathbb{E}\,[Y|c, d, l, m]\,P\,(l|c, d)\,P\,(c)\,, \tag{4.10}$$

where $\mathbb{E}\,[Y|c, d, l, m]$ denotes the expected value of the outcome Y among individuals for whom $C = c$, $D = d$, $L = l$, and $M = m$. Similarly, $P\,(l|c, d)$ denotes the conditional probability that $L = l$ among individuals with $C = c$ and $D = d$, while $P\,(c)$ denotes the marginal probability that $C = c$.

The quantity in Equation 4.10 is obtained by first computing the outcome means conditional on the baseline and exposure-induced confounders, level d of the exposure, and level m of the mediator. These means are then averaged over the distribution of the exposure-induced confounders, given the baseline confounders and exposure to d. The resulting quantities, denoted by $\sum_{l} \mathbb{E}\,[Y|c, d, l, m]\,P\,(l|c, d)$, are then averaged together again, this time over the marginal distribution of the baseline confounders. Under assumptions (d.i) to (d.iv), this yields a quantity equal to $\mathbb{E}\,[Y\,(d, m)]$, the expected value of the joint potential outcomes when the exposure and mediator are set to d and m, respectively.

The nonparametric identification formula contrasts this first quantity with another:

$$\sum_{l,c} \mathbb{E}\left[Y|c,d^*,l,m\right] P\left(l|c,d^*\right) P\left(c\right). \tag{4.11}$$

Here, $\mathbb{E}\left[Y|c,d^*,l,m\right]$ denotes the expected value of the outcome Y among individuals with $C=c$, $D=d^*$, $L=l$, and $M=m$. Similarly, $P\left(l|c,d^*\right)$ denotes the conditional probability that $L=l$ among individuals with $C=c$ and $D=d^*$, while $P\left(c\right)$ is defined as before.

To obtain the quantity in Equation 4.11, we first compute a set of outcome means given the baseline and exposure-induced confounders, level d^* of the exposure, and level m of the mediator. Next, these means are averaged over the distribution of the exposure-induced confounders, given the baseline confounders and exposure to d^*. The resulting quantities, denoted by $\sum_l \mathbb{E}\left[Y|c,d^*,l,m\right] P\left(l|c,d^*\right)$, are then averaged together a final time, now over the marginal distribution of the baseline confounders. When the assumptions outlined previously are satisfied, this yields a quantity equal to $\mathbb{E}\left[Y\left(d^*,m\right)\right]$, the expected value of the joint potential outcomes under exposure to d^* and level m of the mediator.

In substantive terms, this version of the nonparametric identification formula computes the controlled direct effect by perfectly stratifying the target population based on all levels of the baseline confounders. Then, within each stratum, it evaluates the mean of the outcome among those with different levels of the exposure (d versus d^*) but the same levels of the exposure-induced confounders and the mediator. Next, among those exposed to d, it averages these means over the distribution of the exposure-induced confounders observed in this group. Similarly, it also averages the outcome means over the distribution of the exposure-induced confounders among those exposed to d^*. The formula then calculates the difference between these quantities within strata of the baseline confounders and averages them together, weighting each by the relative size of the stratum.

To illustrate, consider the nonparametric identifiability of the $CDE\left(1,0,50\text{K}\right)$ in the NLSY. This effect represents the expected difference in CES-D scores at age 40 if all individuals earned an annual income of \$50,000 between age 35 to 39 but had, versus had not, attended college earlier. It can be nonparametrically identified if the following conditions are met: (d.i) college attendance is conditionally independent of the joint potential outcomes, given the baseline confounders; (d.ii) income is conditionally independent of the joint potential outcomes, given the baseline confounders, college attendance, and unemployment status between age 35 to 39; (d.iii) there is a positive probability of college attendance and each level of income, conditional on their antecedents; and (d.iv) observed scores on the CES-D are consistent with the joint potential outcomes.

In this example, assumption (d.i) requires that there are no unmeasured factors influencing college attendance and depression beyond the variables included among the baseline confounders, such as gender, race, family background, and so on. Similarly, assumption (d.ii) requires that there are no unmeasured factors that jointly affect income and depression, apart from the baseline confounders, college attendance, and unemployment status. Assumption (d.iii) requires that there are individuals who attended college, as well as some who did not, within every subpopulation defined by the baseline confounders. It further requires the existence of individuals at every level of the income distribution, conditional on the baseline confounders, college attendance, and unemployment status. Lastly, assumption (d.iv) essentially stipulates that "income" and "college attendance" must not have multiple forms with heterogeneous effects on depression, and that the influence of income and education must not spillover from one individual to another.

As the NLSY is an observational study, there is no guarantee that these assumptions hold true in our empirical example. In fact, it is likely that they are violated, to some extent. For instance, assumptions (d.i) and (d.ii) may not hold due to the presence of unmeasured factors that could confound the exposure-outcome or mediator-outcome relationship, such as socioemotional skills or disability status. These challenges underscore the difficulty of analyzing causal mediation in observational studies and highlight the utility of joint or sequentially randomized experiments, in which the conditions for identifying controlled direct effects can be met by design. Nevertheless, analyzing controlled direct effects, even in observational studies like the NLSY, can still yield insights into causal mediation, provided that researchers carefully assess potential violations of their identification assumptions and qualify their inferences accordingly.

4.4.2 Nonparametric Identification of Interventional Direct and Indirect Effects

Nonparametric identification of interventional direct and indirect effects requires stronger assumptions compared to those needed for identifying controlled direct effects. However, these assumptions remain considerably weaker than those required for identifying the natural effects decomposition, as they do not include a cross-world independence restriction that is violated whenever exposure-induced confounders are present.

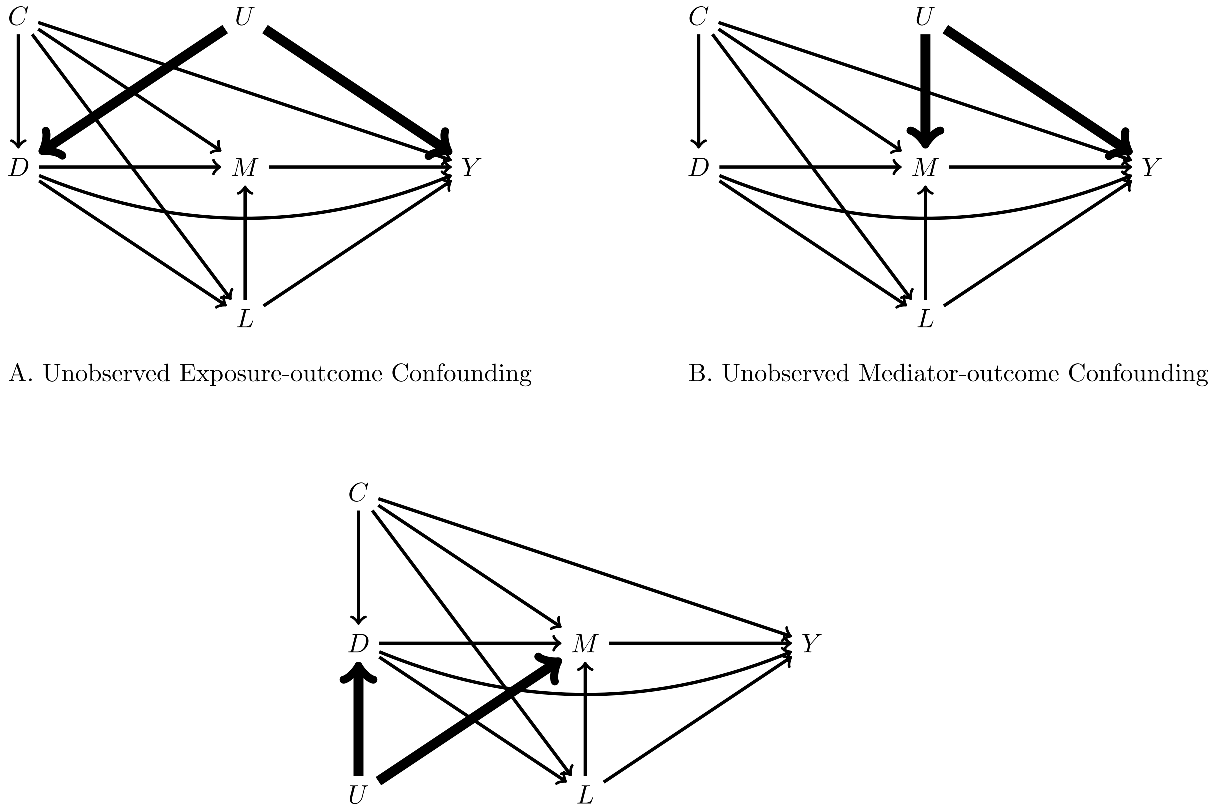
Specifically, nonparametric identification of the $IDE\left(d, d^*\right)$ and $IIE\left(d, d^*\right)$, and by extension, also the $OE\left(d, d^*\right)$, can be achieved under the following set of assumptions: (e.i) conditional independence of the exposure with respect to the joint potential outcomes, (e.ii) conditional independence of the mediator with respect to the joint potential outcomes, (e.iii) conditional independence of the exposure with respect to the potential values for the mediator, (e.iv) sequential positivity of the exposure and mediator, and (e.iv) consistency of the observed and potential values for both the mediator and outcome.

Assumption (e.i). The first of these assumptions requires that the exposure is independent of the joint potential outcomes, conditional on the baseline confounders. Substantively, this implies that there must not be any unobserved exposure-outcome confounders. This assumption is analogous to assumption (d.i) from Section 4.4.1 above. It is also equivalent to assumptions (b.i) and (c.i) from Chapter 3, concerning mediation analysis in the absence of exposure-induced confounders.

Assumption (e.ii). Nonparametric identification of interventional effects also requires that the mediator is independent of the joint potential outcomes, conditional on the baseline confounders, the exposure, and the exposure-induced confounders. In substantive terms, this assumption requires the absence of any unobserved mediator-outcome confounders. It is equivalent to assumption (d.ii) from Section 4.4.1 above. It is also similar to assumptions (b.ii) and (c.ii) from Chapter 3, but it differs from them by additionally incorporating a set of observed exposure-induced confounders that influence the mediator and outcome.

Assumption (e.iii). In addition, nonparametric identification of interventional effects requires that the exposure is independent of the potential values of the mediator, given the baseline confounders. Formally, this assumption can be expressed as $M\left(d\right) \perp D|C$, which is identical to assumption (c.iii) from Section 3.3.3 concerning the natural effects decomposition. Substantively, this assumption implies that there must not be any unobserved confounding of the relationship between the exposure and mediator, after accounting for the baseline confounders.

Assumptions (e.i) to (e.iii) would all be satisfied in data generated from a process resembling Figure 4.3, where no unobserved factors jointly influence both D and Y, M and Y, or D and M. In contrast, Figure 4.6 illustrates scenarios in which each of these assumptions would be violated. Specifically, assumption (e.i) would be violated if the data were generated from a process resembling Panel A of Figure 4.6, where an unobserved variable confounds the exposure-outcome relationship. Similarly, in Panel B, assumption (e.ii) would be violated due to the presence of an unobserved mediator-outcome confounder, while in Panel C, assumption (e.iii) fails because of exposure-mediator confounding by an unobserved variable.



A. Unobserved Exposure-outcome Confounding

B. Unobserved Mediator-outcome Confounding

C. Unobserved Exposure-mediator Confounding

Figure 4.6: Graphical Illustration of Unobserved Exposure-outcome, Mediator-outcome, and Exposure-mediator Confounding.

Note: D denotes the exposure, M denotes a focal mediator of interest, L denotes an exposure-induced confounder, Y denotes the outcome, C denotes a set of baseline confounders, and U denotes a set of unobserved confounders.

Assumption (e.iv). Beyond the conditional independence assumptions outlined previously, nonparametric identification of interventional effects also requires an assumption of sequential positivity. This assumption stipulates that there must be a positive probability of all values for the exposure, conditional on the baseline confounders, and a positive probability of all values for the mediator, given the baseline confounders, the exposure, and the exposure-induced confounders. It is equivalent to assumption (d.iii) from Section 4.4.1 above.

Assumption (e.v). Finally, nonparametric identification of interventional effects depends on a multipart consistency assumption. Specifically, this assumption requires that the observed and potential values of the mediator and outcome must align with each other. It can be formally expressed as follows:

$$Y = Y\left(D, M\right) \text{ and } M = M\left(D\right), \tag{4.12}$$

where D and M denote an individual's observed values on the exposure and mediator, respectively. The first component of this assumption, $Y = Y\left(D, M\right)$, requires that an individual's observed outcome Y must be identical to their joint potential outcome under the levels of the exposure and mediator they did in fact experience. The second component, $M = M\left(D\right)$, further requires that the observed value of the mediator M must align with its potential value under the level of exposure an individual actually experienced.

Nonparametric identification formulas. When assumptions (e.i) to (e.v) are satisfied, the interventional direct effect can be expressed as a function of observable data. This function, termed the nonparametric identification formula for the $IDE\left(d, d^*\right)$, is given by:

$$IDE\left(d, d^*\right) = \sum_{l,m,c} \left(\mathbb{E}\left[Y|c, d, l, m\right] P\left(l|c, d\right) - \mathbb{E}\left[Y|c, d^*, l, m\right] P\left(l|c, d^*\right)\right) P\left(m|c, d^*\right) P\left(c\right), \tag{4.13}$$

where $P\left(m|c, d^*\right)$ denotes the conditional probability that $M = m$ among individuals for whom $C = c$ and $D = d^*$, and all other terms are defined as before. In Appendix H, we provide a step-by-step derivation of this expression, where we also demonstrate how it simplifies to the nonparametric identification formula for the natural direct effect when exposure-induced confounders are absent.

If assumptions (e.i) to (e.v) are met, the interventional indirect effect can also be expressed as a function of observable data. We refer to this function as the nonparametric identification formula for the $IIE\left(d, d^*\right)$, which can be represented as follows:

$$IIE\left(d, d^*\right) = \sum_{l,m,c} \left(P\left(m|c, d\right) - P\left(m|c, d^*\right)\right) \mathbb{E}\left[Y|c, d, l, m\right] P\left(l|c, d\right) P\left(c\right). \tag{4.14}$$

In this expression, $P\left(m|c, d\right)$ is the conditional probability that $M = m$ among individuals with $C = c$ and $D = d$, while all other terms are defined as before. Appendix I contains a step-by-step derivation of this identification formula, and it also shows how Equation 4.14 simplifies to the nonparametric identification formula for the natural indirect effect in the absence of exposure-induced confounding.

Because the overall effect is just obtained by combining interventional direct and indirect effects, it too can be nonparametrically identified under assumptions (e.i) to (e.v). Specifically, when these assumptions are met, the $OE\left(d, d^*\right)$ can be expressed as the following function of observable data:

$$\begin{aligned}OE\left(d,d^{*}\right)=&\sum_{l,m,c}\mathbb{E}\left[Y|c,d,l,m\right]P\left(l|c,d\right)P\left(m|c,d\right)P\left(c\right)\\&-\sum_{l,m,c}\mathbb{E}\left[Y|c,d^{*},l,m\right]P\left(l|c,d^{*}\right)P\left(m|c,d^{*}\right)P\left(c\right),\end{aligned}\tag{4.15}$$

where all terms are defined as before. We refer to this expression as the nonparametric identification formula for the overall effect.

The identification formulas for both the interventional direct and interventional indirect effects share an important expression:

$$\sum_{l,m,c}\mathbb{E}\left[Y|c,d,l,m\right]P\left(l|c,d\right)P\left(m|c,d^{*}\right)P\left(c\right).\tag{4.16}$$

This quantity is obtained by first computing a set of outcome means, conditional on the baseline and exposure-induced confounders, level d of the exposure, and level m of the mediator. Next, among individuals exposed to d and who have the same levels of the baseline confounders, these means are averaged over the distribution of the exposure-induced confounders. The resulting quantities, denoted by $\sum_{l}\mathbb{E}\left[Y|c,d,l,m\right]P\left(l|c,d\right)$, are then averaged together again, this time over the distribution of the mediator given the baseline confounders and the other exposure level, d^{*}. Finally, the quantities resulting from these calculations, denoted by $\sum_{l,m}\mathbb{E}\left[Y|c,d,l,m\right]P\left(l|c,d\right)P\left(m|c,d^{*}\right)$, are averaged together a final time, now over the marginal distribution of the baseline confounders. When assumptions (e.i) to (e.v) hold, this yields a quantity equal to $\mathbb{E}\left[Y\left(d,\mathcal{M}\left(d^{*}|C\right)\right)\right]$, the expected value of the randomized potential outcomes where the exposure is set to one level, d, but the mediator is selected randomly from its distribution under the alternative exposure, d^{*}.

In the nonparametric identification formula for the $IDE\left(d,d^{*}\right)$, this first quantity is contrasted with another:

$$\sum_{l,m,c}\mathbb{E}\left[Y|c,d^{*},l,m\right]P\left(l|c,d^{*}\right)P\left(m|c,d^{*}\right)P\left(c\right).\tag{4.17}$$

To obtain this second quantity, we first compute a set of outcome means, conditional on the baseline and exposure-induced confounders, level d^{*} of the exposure, and level m of the mediator. These means are then averaged over the distribution of the exposure-induced confounders, given the baseline confounders and exposure to d^{*}. The resulting quantities, denoted by $\sum_{l}\mathbb{E}\left[Y|c,d^{*},l,m\right]P\left(l|c,d^{*}\right)$, are then averaged together again, now over the distribution of the mediator given the baseline confounders and exposure to d^{*}. Finally, the quantities resulting from these calculations, denoted by $\sum_{l,m}\mathbb{E}\left[Y|c,d^{*},l,m\right]P\left(l|c,d^{*}\right)P\left(m|c,d^{*}\right)$, are averaged one last time over the marginal distribution of the baseline confounders. If the assumptions outlined previously are met, this yields a quantity equal to $\mathbb{E}\left[Y\left(d^{*},\mathcal{M}\left(d^{*}|C\right)\right)\right]$, which represents the expected value of the randomized potential outcomes where the exposure is set to d^{*} and the mediator is randomly selected from its distribution under this same level of the exposure.

In the nonparametric identification formula for the $IIE\left(d,d^{*}\right)$, however, the quantity in Equation 4.16 is contrasted with a different function of observable data. Specifically, it is contrasted with:

$$\sum_{l,m,c}\mathbb{E}\left[Y|c,d,l,m\right]P\left(l|c,d\right)P\left(m|c,d\right)P\left(c\right).\tag{4.18}$$

This third quantity is computed by obtaining a set of outcome means, conditional on the baseline and exposure-induced confounders, level d of the exposure, and level m of the mediator. These means are then averaged over the distribution of the exposure-induced confounders, given the baseline confounders and exposure to d. As before, the resulting quantities are denoted by $\sum_{l} \mathbb{E}\left[Y|c,d,l,m\right] P\left(l|c,d\right)$, and they are then averaged together again, but now over the distribution of the mediator among individuals exposed to d and who have the same levels on the baseline confounders. Lastly, the quantities resulting from these calculations, denoted by $\sum_{l,m} \mathbb{E}\left[Y|c,d,l,m\right] P\left(l|c,d\right) P\left(m|c,d\right)$, are averaged together a final time over the marginal distribution of the baseline confounders. Under assumptions (e.i) to (e.v), this yields a quantity equal to $\mathbb{E}\left[Y\left(d,\mathcal{M}\left(d|C\right)\right)\right]$, representing the expected value of the randomized potential outcomes where the exposure is set to d and the mediator is randomly selected from its distribution under this same level of the exposure.

Note that, in the nonparametric identification formula for the $IDE\left(d,d^{*}\right)$, we average the conditional means of the outcome over two different distributions of the exposure-induced confounder L–specifically, $P\left(l|c,d\right)$ and $P\left(l|c,d^{*}\right)$–but we only average these means over a single distribution of the focal mediator, given by $P\left(m|c,d^{*}\right)$. This approach adjusts for the confounding influence of L on the $M \rightarrow Y$ relationship while including the mediating influence of L along the $D \rightarrow L \rightarrow Y$ path as part of the $IDE\left(d,d^{*}\right)$. Conversely, in the nonparametric identification formula for the $IIE\left(d,d^{*}\right)$, we average a conditional mean of the outcome over two different distributions of the focal mediator, denoted by $P\left(m|c,d\right)$ and $P\left(m|c,d^{*}\right)$, but we only average this mean over a single distribution of the exposure-induced confounder L, given by $P\left(l|c,d\right)$. This approach adjusts for the confounding influence of L while excluding its mediating influence from the $IIE\left(d,d^{*}\right)$. Essentially, by averaging over the distributions of L and M in different ways, these identification formulas partition the mediating influence of L into the interventional direct effect and the mediating influence of M into the interventional indirect effect, all while adjusting appropriately for exposure-induced confounding.

The nonparametric identification formula for the $OE\left(d,d^{*}\right)$ is given by the sum of the identification formulas for the $IDE\left(d,d^{*}\right)$ and $IIE\left(d,d^{*}\right)$. It therefore contrasts the quantity in Equation 4.18 with that in Equation 4.17. As outlined previously, the first of these quantities is equal to $\mathbb{E}\left[Y\left(d,\mathcal{M}\left(d|C\right)\right)\right]$, while the second is equal to $\mathbb{E}\left[Y\left(d^{*},\mathcal{M}\left(d^{*}|C\right)\right)\right]$, provided that assumptions (e.i) to (e.v) are met.

In substantive terms, these identification formulas compute interventional effects through a process of stratification, averaging, and weighting. The target population is first perfectly stratified by all levels of the baseline and exposure-induced confounders, the exposure itself, and the mediator. Within each stratum, the mean of the outcome is evaluated, and then these means are then averaged together in different ways. Initially, they are averaged across levels of the exposure-induced confounders and the mediator, all within levels of the baseline confounders. When computing these averages, the stratum-specific means are given different weights, which come from the probability of experiencing different levels of the exposure-induced confounders and the mediator. After combining the outcome means across levels of the exposure-induced confounders and the mediator, the resulting quantities are averaged together again, now across levels of the baseline confounders. When computing these averages, the weights are given by the probability of experiencing different levels of the baseline confounders. Finally, the effects of interest are computed by comparing these different weighted means.

Nonparametric identification of interventional effects relies on strong assumptions. Nevertheless, these assumptions are less stringent than those required for identifying natural direct and indirect effects, as they allow for exposure-induced confounding. Moreover, unlike the natural effects decomposition, interventional effects can be identified by experimental design, at least in theory.

To illustrate, imagine a multi-arm experiment where participants are randomly assigned to the different arms of the study. In the first arm, participants are assigned to exposure d, and then their mediator is measured under this exposure. In the second arm, participants are assigned to exposure d^*, and their mediator is measured under this alternative exposure. The mediator values observed in the first arm comprise the distribution from which $\mathcal{M}(d|C)$ will be randomly selected, while the second arm of this experiment generates the distribution from which $\mathcal{M}(d^*|C)$ will be drawn.

These two distributions inform the subsequent arms of the experiment. Specifically, in the third arm, participants are assigned to exposure d, and their mediator is randomly selected from the distribution observed in the first arm. Measures of the outcome for each participant would then represent the randomized potential outcome $Y(d, \mathcal{M}(d|C))$.

Moving on to the fourth arm of the experiment, participants are here assigned to exposure d^*, and their mediator is randomly selected from its distribution in the second arm. Outcome measures obtained from participants in this arm would represent $Y(d^*, \mathcal{M}(d^*|C))$.

Finally, in the fifth arm, participants are assigned to exposure d, and their mediator is randomly selected from its distribution in the second arm of the experiment. Measures of the outcome for each participant in this arm would represent $Y(d, \mathcal{M}(d^*|C))$. The interventional effects of interest can then be computed by comparing means of the observed outcomes across the third, fourth, and fifth arms of the experiment.

Under this design, random assignment of participants to different arms would eliminate exposure-outcome and exposure-mediator confounding, while random assignment of participants to different values of the mediator within arms would remove any mediator-outcome confounding. Although such an experiment may not be logistically or ethically feasible, depending on the details of any given application, it is at least conceivable in certain contexts, and thus experimental identification of interventional effects remains a possibility. In contrast, identifying natural direct and indirect effects cannot be achieved by experimental design alone. We revisit these complexities in Chapter 7.

Now consider the nonparametric identifiability of the $IDE(1,0)$, $IIE(1,0)$, and $OE(1,0)$ in the NLSY. Together, these effects describe how college attendance influences depression through a causal process involving income. They can be nonparametrically identified if the following conditions are met: (e.i) college attendance is conditionally independent of the joint potential outcomes, given the baseline confounders; (e.ii) income is conditionally independent of the joint potential outcomes, given the baseline confounders, college attendance, and unemployment; (e.iii) college attendance is conditionally independent of the potential values of the mediator, given the baseline confounders; (e.iv) there is a positive probability of attending college and earning at all income levels, conditional on prior variables; and (e.v) the observed and potential values of the mediator and outcome are consistent with each other.

In the NLSY, assumptions (e.i) to (e.iii) require the absence of unobserved factors that confound the relationships of college attendance with depression, income with depression, and college attendance with income. Furthermore, assumption (e.iv) implies that individuals must have the potential to experience each different level of education and income within subpopulations defined by the antecedents to these variables. And assumption (e.v) dictates that there must not be multiple versions of education and income with heterogeneous effects on depression or any interference between different individuals.

These assumptions present challenges in our analysis of the NLSY, given the existence of unobserved variables that could influence education, income, and depression simultaneously. The exposure–college attendance–also encompasses a diverse set of experiences with potentially variable effects on mental health. Similarly, not all forms of the mediator are created equal. For example, it's possible that an additional \$10,000 in earned wages might have a different impact on mental health compared to receiving \$10,000 in lottery winnings, raising questions about the consistency assumption.

Satisfying these assumptions exactly in any observational study is difficult, and even in an experimental setting, it requires a complex design that may be practically challenging to execute. Thus, analyzing causal mediation remains a formidable task no matter the research design or target estimands, which necessitates cautious and appropriately qualified inferences at all times.

4.5 Nonparametric Estimation

In this section, we shift our attention from nonparametric identification with full population data to nonparametric estimation using a random sample. Specifically, we explain how to construct consistent estimators for interventional and controlled direct effects in the presence of exposure-induced confounding, without imposing any functional form restrictions on the probability distribution from which the data were sampled. As outlined previously, a consistent estimator is one that converges to its target estimand as the sample size increases indefinitely, or in other words, it becomes more accurate when fed more data.

Nonparametric estimation can be challenging or even impossible to implement in practice due to the twin problems of data sparsity and the curse of dimensionality. To circumvent these challenges, we use a contrived example with only a few simplified measures. Nevertheless, nonparametric estimation remains relevant because it provides consistent estimates under weaker assumptions than parametric approaches, whenever it can be practically implemented with available data.

Nonparametric estimation of interventional effects essentially just involves substituting population quantities with their sample analogues in the identification formulas from Section 4.4. To illustrate, consider the sample data presented in Table 4.1 from the NLSY. This table summarizes case counts and sample means, separately by maternal education C, college attendance D, unemployment status L, and income M.

Specifically, $\bar{Y}_{c,d,l,m}$ in this table denotes the sample mean of the standardized CES-D scores among respondents for whom $C = c$, $D = d$, $L = l$, and $M = m$, while $n_{c,d,l,m}$ represents the number of respondents with these values on each of the covariates. To simplify our analysis and enable nonparametric estimation with the NLSY, the mediator of interest–income–has been recast as a binary variable, indicating whether a respondent earned over \$50,000 annually from age 35 to 39.

Based on these data, a nonparametric estimator for the controlled direct effect can be expressed as follows:

$$\begin{aligned}\widehat{CDE}\left(d,d^*,m\right)^{npl} &= \sum_{l,c}\left(\hat{\mathbb{E}}\left[Y|c,d,l,m\right]\hat{P}\left(l|c,d\right) - \hat{\mathbb{E}}\left[Y|c,d^*,l,m\right]\hat{P}\left(l|c,d^*\right)\right)\hat{P}\left(c\right) \\ &= \sum_{l,c}\left(\bar{Y}_{c,d,l,m}\hat{\pi}_{l|c,d} - \bar{Y}_{c,d^*,l,m}\hat{\pi}_{l|c,d^*}\right)\hat{\pi}_c. \end{aligned} \tag{4.19}$$

In this expression, $\hat{\pi}_c = \sum_{m,l,d} n_{c,d,l,m}/n$ denotes the proportion of sample members for whom $C = c$. Similarly, $\hat{\pi}_{l|c,d} = \sum_m n_{c,d,l,m}/\sum_{m,l} n_{c,d,l,m}$ denotes the proportion of sample members for whom $L = l$ among those with $C = c$ and $D = d$, while $\hat{\pi}_{l|c,d^*} = \sum_m n_{c,d^*,l,m}/\sum_{m,l} n_{c,d^*,l,m}$ is defined analogously. As in the previous chapter, we use "hats" to distinguish estimators from estimands and superscripts to differentiate among distinct estimators. The "npl" superscript in Equation 4.19 is intended to signify that it is a nonparametric estimator for the controlled direct effect in the presence of exposure-induced confounding.

Applying this expression to the data in Table 4.1, an estimate for the controlled direct effect of college attendance on depression, if individuals had all earned more than \$50,000, is given by $\widehat{CDE}\left(1,0,50\text{K}\right)^{npl} = -0.17$. This estimate suggests that attending college would still reduce CES-D scores by 0.17 standard

Table 4.1: Case Counts and Sample Means for CES-D Scores (Y) by College Attendance (D), Unemployment Status (L), Household Income (M), and Maternal Education (C), NLSY.

			Maternal Education			
Respondent Education	Unemployment Status	Household Income	No College ($C = 0$)		Attended College ($C = 1$)	
			$\bar{Y}_{c,d,l,m}$	$n_{c,d,l,m}$	$\bar{Y}_{c,d,l,m}$	$n_{c,d,l,m}$
No College	Never Unemp.	$<$ \$50K ($M = 0$)	.07	1219	.05	98
($D = 0$)	($L = 0$)	$\geq$ \$50K ($M = 1$)	$-.17$	528	$-.11$	74
	Ever Unemp.	$<$ \$50K ($M = 0$)	.35	460	.39	36
	($L = 1$)	$\geq$ \$50K ($M = 1$)	.04	60	.34	3
Attended Col.	Never Unemp.	$<$ \$50K ($M = 0$)	$-.07$	211	.03	85
($D = 1$)	($L = 0$)	$\geq$ \$50K ($M = 1$)	$-.30$	358	$-.22$	247
	Ever Unemp.	$<$ \$50K ($M = 0$)	.24	56	.14	18
	($L = 1$)	$\geq$ \$50K ($M = 1$)	$-.16$	36	$-.23$	24

Note: CES-D scores have been standardized to have zero mean and unit variance. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch4/table_4-1`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

deviations, on average, even if everyone earned at least a “middle class” income. For brevity, we have refrained from illustrating the calculations step by step, as in the previous chapter. Instead, we leave this as an exercise for the reader.

Similarly, a nonparametric estimator for the interventional direct effect can be expressed as follows:

$$\begin{aligned}\widehat{IDE}\left(d,d^{*}\right)^{npl} &= \sum_{l,m,c}\Big(\hat{\mathbb{E}}\left[Y|c,d,l,m\right]\hat{P}\left(l|c,d\right)-\hat{\mathbb{E}}\left[Y|c,d^{*},l,m\right]\hat{P}\left(l|c,d^{*}\right)\Big)\hat{P}\left(m|c,d^{*}\right)\hat{P}\left(c\right)\\ &= \sum_{l,m,c}\left(\bar{Y}_{c,d,l,m}\hat{\pi}_{l|c,d}-\bar{Y}_{c,d^{*},l,m}\hat{\pi}_{l|c,d^{*}}\right)\hat{\pi}_{m|c,d^{*}}\hat{\pi}_{c}, \qquad (4.20)\end{aligned}$$

where $\hat{\pi}_{m|c,d^{*}}=\sum_{l}n_{c,d^{*},l,m}/\sum_{m,l}n_{c,d^{*},l,m}$ is the proportion of sample members for whom $M = m$ among those with $C = c$ and $D = d^{*}$, and all other terms are defined as before. Applying this expression to the data in Table 4.1 yields an estimate for the interventional direct effect of $\widehat{IDE}\left(1,0\right)^{npl} = -0.16$, which suggests that attending college would reduce CES-D scores by about one-sixth of a standard deviation, on average, even if everyone earned an income selected randomly from the distribution that would have arisen had they not attended college.

A nonparametric estimator for the interventional indirect effect is given by the following expression:

$$\begin{aligned}\widehat{IIE}\left(d,d^{*}\right)^{npl} &= \sum_{l,m,c}\Big(\hat{P}\left(m|c,d\right)-\hat{P}\left(m|c,d^{*}\right)\Big)\hat{\mathbb{E}}\left[Y|c,d,l,m\right]\hat{P}\left(l|c,d\right)\hat{P}\left(c\right)\\ &= \sum_{l,m,c}\left(\hat{\pi}_{m|c,d}-\hat{\pi}_{m|c,d^{*}}\right)\bar{Y}_{c,d,l,m}\hat{\pi}_{l|c,d}\hat{\pi}_{c}, \qquad (4.21)\end{aligned}$$

where $\hat{\pi}_{m|c,d}$ is the proportion of sample members for whom $M = m$ among those with $C = c$ and $D = d$, and all other terms are defined as before. With the data in Table 4.1, applying this expression gives an estimate for the interventional indirect effect of $\widehat{IIE}\left(d, d^*\right)^{npl} = -0.09$. This estimate suggests that CES-D scores would decline by roughly one-tenth of a standard deviation, on average, if individuals earned an income randomly selected from the distribution that would have arisen had they attended college, as opposed to the distribution that would have emerged had they not attended college.

Finally, a nonparametric estimator for the overall effect can be expressed as a sum of the estimators for the interventional direct and indirect effects:

$$\begin{aligned}\widehat{OE}\left(d, d^*\right)^{npl} &= \widehat{IDE}\left(d, d^*\right)^{npl} + \widehat{IIE}\left(d, d^*\right)^{npl} \\ &= \sum_{l,m,c}\left(\bar{Y}_{c,d,l,m}\hat{\pi}_{l|c,d}\hat{\pi}_{m|c,d} - \bar{Y}_{c,d^*,l,m}\hat{\pi}_{l|c,d^*}\hat{\pi}_{m|c,d^*}\right)\hat{\pi}_c. \qquad (4.22)\end{aligned}$$

Applying this expression to the data in Table 4.1 yields an estimate of $\widehat{OE}\left(1, 0\right)^{npl} = -0.24$, which indicates that college attendance reduces CES-D scores by about one-quarter of a standard deviation overall. Taken altogether, then, our results suggest that attending college reduces depression at midlife and that a nontrivial part of this effect is mediated by differences in the distribution of income.

These estimators, however, are only consistent if the assumptions required to nonparametrically identify their target estimands are satisfied. Otherwise, they risk being biased and inconsistent. In our simplified illustration with the NLSY, at least some of these assumptions are likely violated. This is primarily because we only adjusted for a single covariate at baseline, when a multitude of other factors may jointly influence the exposure, mediator, and outcome, resulting in bias due to unobserved confounding.

Furthermore, nonparametric estimation is typically complicated by sparse data and the curse of dimensionality, as discussed extensively in Chapter 3. These challenges stem from our reliance on finite samples that contain many variables or include measures with many values, and they are frequently severe enough to preclude nonparametric estimation entirely. Indeed, this approach would not have been possible to implement with the NLSY had we not artificially simplified our analysis to reduce the dimension of the data and thereby mitigate the problem of sparsity.

Thus, despite its theoretical promise of providing consistent results under weaker assumptions than other methods, nonparametric estimation may be impractical or infeasible in many social science applications. In these instances, estimation strategies that rely on parametric models offer significant advantages, albeit at the cost of more stringent assumptions about the probability distribution from which the data were sampled. In the next section, we introduce a series of parametric estimators for interventional effects, which are more broadly applicable in practice.

4.6 Parametric Estimation

In this section, we discuss how to estimate interventional and controlled direct effects using parametric models. These models place restrictions on the joint distribution of the observed data, allowing parametric estimators to mitigate the problems of sparsity and high dimensionality that often hinder nonparametric methods.

However, the accuracy of parametric estimation depends on avoiding model misspecification. If the underlying models used to construct these estimators are misspecified, they will suffer from systematic bias and fail to converge to their target estimands as the sample size increases, even if our identification assumptions about the absence of unobserved confounding are all met.

Because parametric models are almost always misspecified to some degree, the practical aim when analyzing mediation is to find models that are nearly correct and minimize estimation error. This often requires navigating a trade-off between bias and variance, where complex models with many parameters afford greater protection against bias due to misspecification but yield estimates with higher sampling variability.

Throughout the rest of this section, we present a series of parametric estimators for interventional and controlled direct effects in the presence of exposure-induced confounding. First, we present a regression-based estimator that uses linear models for the mediator, exposure-induced confounders, and outcome. This approach can accommodate both discrete and continuous exposures but performs best when the mediator and outcome are continuous. Next, we introduce a simulation-based estimator that supports a wide range of linear or nonlinear models for these variables. This approach is quite flexible and can be implemented with many different types of exposures, mediators, and outcomes, whether continuous or discrete. Lastly, we discuss estimators based on inverse probability weights, which are derived from models for the exposure, mediator, and exposure-induced confounders. This approach is best suited for applications where all these variables are binary or have a limited number of discrete values.

4.6.1 Estimation using Linear Models: Regression-with-residuals

In this section, we explain how to estimate interventional and controlled direct effects with linear models, using a technique known as *regression-with-residuals* (RWR; Almirall et al. 2010; Wodtke and Almirall 2017; Wodtke 2020; Wodtke et al. 2020; Wodtke and Zhou 2020; Zhou and Wodtke 2019). This approach involves only a slight modification of the regression-based methods we covered in Chapter 3. In that chapter, we assumed there was no exposure-induced confounding and focused on estimating the natural effects decomposition using linear models for the mediator and outcome. Here, we adapt these methods to account for exposure-induced confounding and focus on estimating interventional effects.

The most basic implementation of RWR proceeds as follows. First, we fit a linear model for the mediator, using the exposure and baseline confounders as predictors. Second, for each of the exposure-induced confounders, we also fit a linear model, again using the exposure and baseline confounders as predictors, and then we compute residual terms from these models. Lastly, we fit another linear model, where the outcome is regressed on the exposure, mediator, baseline confounders, and the residual terms obtained in the previous step. All these models are fit using the method of least squares, and estimates for the effects of interest are given by simple functions of their coefficients.

Specifically, RWR estimates for interventional and controlled direct effects can be constructed from the following set of linear models. The first model is for the conditional mean of the mediator, given the exposure and baseline confounders. It can be expressed as follows:

$$\mathbb{E}\left[M|c,d\right] = \beta_0 + \beta_1^T c^{\perp} + \beta_2 d, \tag{4.23}$$

where $c^{\perp} = c - \bar{C}$. This model closely resembles a standard linear regression for the mediator, except that the baseline confounders are centered around their sample means.

The next set of models is for the conditional mean of each exposure-induced confounder L, given the exposure D and baseline confounders C. These models can be expressed as follows:

$$\mathbb{E}\left[L|c,d\right] = \lambda_0 + \lambda_1^T c^{\perp} + \lambda_2 d, \tag{4.24}$$

which represents a standard linear regression for L with the exposure and mean-centered confounders at baseline as predictors. The regressions for L are used to compute residual terms, denoted by $l^{\perp} = l - \mathbb{E}\left[L|c,d\right]$. Because residuals are orthogonal to the predictors in a linear model by design (Fox 2015), $l^{\perp}$ represents a transformation of the exposure-induced confounders that purges them of their association with the exposure and baseline confounders.

The last model is for the conditional mean of the outcome, given the exposure, mediator, baseline confounders, and exposure-induced confounders. It can be expressed as follows:

$$\mathbb{E}\left[Y|c,d,l,m\right] = \gamma_0 + \gamma_1^T c^{\perp} + \gamma_2 d + m\left(\gamma_3 + \gamma_4 d\right) + \gamma_5^T l^{\perp}, \tag{4.25}$$

where $l^{\perp} = l - \mathbb{E}\left[L|c,d\right]$ again denotes the residuals obtained in the previous step. This model is similar to a standard linear regression for the outcome, except that the baseline confounders are centered around their sample means and, in addition, the exposure-induced confounders are centered around their conditional means given the exposure and baseline confounders. In other words, the outcome regression here includes a set of residual terms for the exposure-induced confounders as predictors, rather than the original versions of these variables themselves.

If assumptions (d.i) to (d.iv) hold, and if Equations 4.23 to 4.24 are correctly specified, the controlled direct effect can be expressed as $CDE\left(d,d^*,m\right) = \left(\gamma_2 + \gamma_4 m\right)\left(d - d^*\right)$. Similarly, if assumptions (e.i) to to (e.v) are met and the same set of models are correctly specified, the interventional direct and indirect effects can be expressed as $IDE\left(d,d^*\right) = \left(\gamma_2 + \gamma_4\left(\beta_0 + \beta_2 d^*\right)\right)\left(d - d^*\right)$ and $IIE\left(d,d^*\right) = \beta_2\left(\gamma_3 + \gamma_4 d\right)\left(d - d^*\right)$, respectively, with the overall effect given by their sum.

It is worth noting that these expressions are identical to those for the natural effects decomposition under linear models for the mediator and outcome, as outlined in Section 3.5.1 of the previous chapter. The only difference is that the expressions for the interventional effects are based on an outcome model that additionally adjusts for exposure-induced confounders after these variables have first been residualized.

To summarize, RWR estimation of interventional and controlled direct effects can be implemented through the following steps:

1. **Center the baseline confounders.** For each baseline confounder, compute $C^{\perp} = C - \overline{C}$, which centers these variables around their sample means.

2. **Residualize the exposure-induced confounders.** For each exposure-induced confounder, compute $L^{\perp} = L - \hat{\mathbb{E}}\left[L|C,D\right]$ by fitting a least squares regression of L on C and D and then extracting the residuals.

3. **Fit a linear model for the mediator.** Using $C^{\perp}$ from step 1, compute least squares estimates of Equation (4.23), which can be expressed as

$$\hat{\mathbb{E}}\left[M|C,D\right] = \hat{\beta}_0 + \hat{\beta}_1^T C^{\perp} + \hat{\beta}_2 D.$$

4. **Fit a linear model for the outcome.** Using $C^{\perp}$ from step 1 and $L^{\perp}$ from step 2, compute least

squares estimates of Equation (4.25), which can be expressed as

$$\hat{\mathbb{E}}\left[Y|C,D,L,M\right] = \hat{\gamma}_0 + \hat{\gamma}_1^T C^{\perp} + \hat{\gamma}_2 D + M\left(\hat{\gamma}_3 + \gamma_4 D\right) + \hat{\gamma}_5^T L^{\perp}.$$

5. **Construct effect estimates.** Estimators for the interventional and controlled direct effects of interest are based on functions of coefficients in the fitted models from steps 3 and 4. Specifically, they are given by the following expressions:

$$\begin{aligned} \widehat{CDE}\left(d,d^*,m\right)^{rwr} &= \left(\hat{\gamma}_2 + \hat{\gamma}_4 m\right)\left(d-d^*\right) \\ \widehat{IDE}\left(d,d^*\right)^{rwr} &= \left(\hat{\gamma}_2 + \hat{\gamma}_4\left(\hat{\beta}_0 + \hat{\beta}_2 d^*\right)\right)\left(d-d^*\right) \\ \widehat{IIE}\left(d,d^*\right)^{rwr} &= \hat{\beta}_2\left(\hat{\gamma}_3 + \hat{\gamma}_4 d\right)\left(d-d^*\right) \\ \widehat{OE}\left(d,d^*\right)^{rwr} &= \left(\hat{\gamma}_2 + \hat{\gamma}_4\left(\hat{\beta}_0 + \hat{\beta}_2 d^*\right) + \hat{\beta}_2\left(\hat{\gamma}_3 + \hat{\gamma}_4 d\right)\right)\left(d-d^*\right). \end{aligned} \tag{4.26}$$

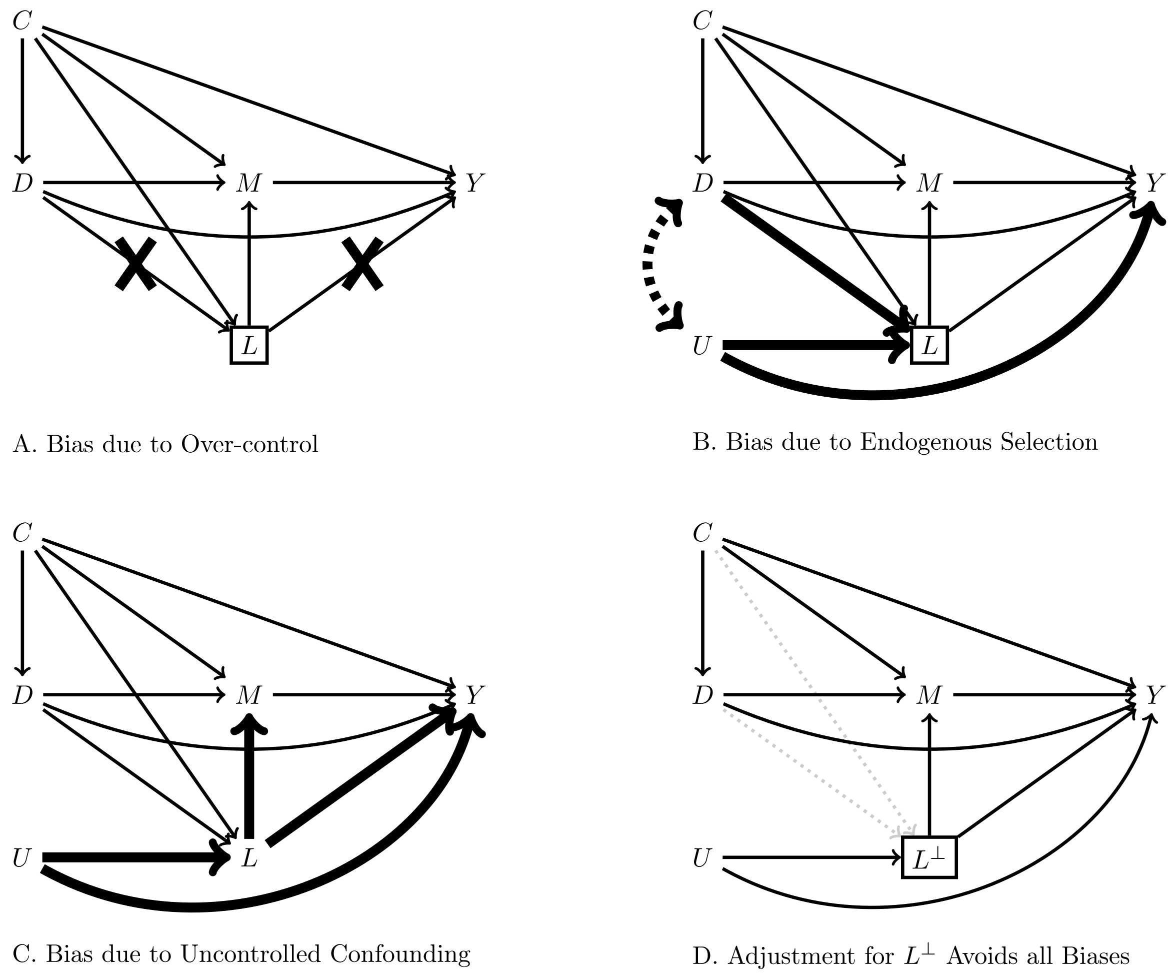
   Here, the "rwr" superscript indicates that these expressions are regression-with-residuals estimators.

This procedure yields consistent estimates as long as our identification assumptions hold and our regression models for the outcome, mediator, and exposure-induced confounders are correctly specified.

The defining feature of RWR is that it adjusts for a residual transformation of the exposure-induced confounders. Why adjust for these residual terms rather than the exposure-induced confounders themselves? When adjusting for exposure-induced confounders, naively including these variables as predictors in an outcome regression can introduce bias, even if the effects of interest are identifiable from the observed data. Conversely, failing to make appropriate adjustments for the exposure-induced confounders also leads to bias. As a result, exposure-induced confounders seemingly pose a "damned if you do and damned if you don't" dilemma with regard to covariate adjustment. However, residualizing these variables before including them in an outcome regression circumvents these problems entirely (Wodtke and Almirall 2017; Wodtke 2020; Wodtke and Zhou 2020; Wodtke et al. 2020; Zhou and Wodtke 2019).

To appreciate this, consider the graph in Panel A of Figure 4.7. Recall that a path in a DAG is "blocked" when it contains either (i) an outcome of two or more variables, known as a collider, that has not been conditioned upon, or (ii) a non-collider that has been conditioned upon. Otherwise, a path is "unblocked," and the variables it connects are statistically associated (Elwert 2013; Pearl 2009). The graph in Panel A shows that adjusting for the exposure-induced confounder L blocks the causal path $D \rightarrow L \rightarrow Y$, which would lead to bias in estimates of interventional and controlled direct effects due to *over-control of intermediate pathways.* In other words, adjusting for L would "control away" a portion of the causal association between the exposure and outcome–specifically, the part that operates through the exposure-induced confounder–which is a component of the interventional and controlled direct effects of interest.

Next, consider the graph in Panel B of Figure 4.7, which additionally includes an unobserved variable U that directly affects both the confounder L and outcome Y but not the exposure D or mediator of interest M. This graph shows that adjusting for L unblocks the non-causal path $D \rightarrow L \leftarrow U \rightarrow Y$ from the exposure to the outcome, thereby introducing bias due to *endogenous selection* (Elwert and Winship 2014). Specifically, the graph shows that L is a collider of D and U, so adjusting for L would induce a non-causal association between the exposure and the unobserved variable, as indicated by the dashed bidirectional arrow connecting them. Moreover, because U also affects Y, adjusting for L would induce a non-causal association between the exposure and outcome as well. This spurious association between the exposure D and outcome Y also



A. Bias due to Over-control

B. Bias due to Endogenous Selection

C. Bias due to Uncontrolled Confounding

D. Adjustment for $L^{\perp}$ Avoids all Biases

Figure 4.7: Graphical Illustration of Bias due to Over-control, Endogenous Selection, and Uncontrolled Confounding.

Note: D denotes the exposure, M denotes a focal mediator of interest, L denotes an exposure-induced confounder, Y denotes the outcome, C denotes a set of baseline confounders, and U denotes an unobserved variable. In addition, $L^{\perp} = L - \hat{\mathbb{E}}(L|C, D)$ denotes an exposure-induced confounder that has been residualized with respect to the exposure D and baseline confounders C. A box around a variable is used to denote that it has been conditioned upon (i.e., “controlled” or “adjusted for”).

leads to bias in estimates of interventional and controlled direct effects.

Lastly, Panel C of Figure 4.7 illustrates that, when L is not conditioned upon, several non-causal pathways from the mediator to the outcome (specifically, $M \leftarrow L \rightarrow Y$ and $M \leftarrow L \leftarrow U \rightarrow Y$) remain unblocked. This leaves a non-causal association between the mediator M and outcome Y undisturbed, leading to bias in estimates of interventional effects due to uncontrolled confounding. Thus, our effect estimates appear to suffer from bias regardless of how we handle the exposure-induced confounders: adjusting for these variables could lead to bias from over-control and endogenous selection, while not adjusting for them results in bias due to mediator-outcome confounding.

RWR circumvents these problems by adjusting for a residual transformation of the exposure-induced confounders, as illustrated in Panel D of Figure 4.7. This graph shows how residualizing the exposure-induced confounders with respect to the exposure and baseline confounders effectively neutralizes the causal paths emanating from D and C into L. As a result, the residualized confounders can be included in an outcome regression to adjust for mediator-outcome confounding, while sidestepping the pitfalls associated with naive adjustment for variables influenced by the exposure. Essentially, RWR avoids bias due to over-control and endogenous selection because the residual terms, denoted by $L^{\perp}$, are no longer related to the exposure D, and it avoids confounding bias because adjusting for $L^{\perp}$ together with D and C in a regression for the outcome Y sufficiently controls for mediator-outcome confounding.

With RWR, the outcome model must be linear in the parameters, and thus it is best suited for applications where Y is continuous. RWR can also be used when the outcome is binary, ordinal, or counts, as long as a linear model can be reasonably assumed to approximate its true but unknown conditional expected value of the outcome in the specific application at hand.

Although the linearity requirement for the outcome model is restrictive, RWR is flexible in other ways. For example, it can readily incorporate two-way interactions between $C^{\perp}$ and D, $C^{\perp}$ and M, or $L^{\perp}$ and M, which allow the effects of the exposure and mediator to vary across levels of the confounders. As long as these interaction terms are constructed with the mean-centered and residualized transformations of the confounders, computing the interventional and controlled direct effects of interest proceeds exactly as outlined previously.

RWR offers additional flexibility in terms of the choice of models for the exposure-induced confounders. Although we introduced the method using linear regressions for these variables, the exposure-induced confounders can be residualized using any generalized linear model (GLM) that is appropriate for their level of measurement. For example, if L is binary, it could be modeled and residualized using a logistic or probit regression. A convenient feature of RWR is that its parametric expressions for the effects of interest, as given in Equation (4.26), are insensitive to the choice of models for the exposure-induced confounders. Thus, regardless of the models used to residualize L, RWR estimation of interventional and controlled direct effects follows the same steps as outlined above.

It is also possible to implement RWR using a mediator model in the broader family of GLMs instead of solely relying on linear regression. However, when RWR is implemented with a nonlinear model for the mediator, constructing consistent estimates for the interventional effects of interest necessitates different and more intricate functions of model coefficients, compared to the relatively simple expressions provided in Equation (4.26) above. We limit our discussion to implementations of RWR using a linear model for the mediator, but the method can be extended even further to incorporate nonlinear models for $\mathbb{E}[M|c, d]$ at the cost of some added complexity (Wodtke and Zhou 2020).

Using data from the NLSY, Table 4.2 presents point estimates for the interventional and controlled direct

Table 4.2: Interventional Effects of College Attendance on CES-D Scores as Estimated from the NLSY using Regression-with-residuals.

Estimand	Point Estimates		
	RWR with $D \times M$ Interaction	RWR with $\ln(M)$ and $D \times \ln(M)$ Interaction	RWR with $\ln(M)$ and All Two-way Interactions
$OE(1,0)$	−.080	−.076	−.112
$IDE(1,0)$	−.061	−.040	−.041
$IIE(1,0)$	−.019	−.036	−.071
$CDE(1,0,50\text{K})$	−.052	−.027	−.084

Note: Estimates are expressed in standard deviation units. D denotes college attendance, M denotes income, and $\ln(M)$ denotes the natural log of income. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch4/table_4-2`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

effects of college attendance on depression, as mediated by income. These estimates were computed using RWR with linear models for the mediator, outcome, and exposure-induced confounder (unemployment), all of which adjust for the full set of baseline confounders, including measures of race, gender, parental education, and so on.

The estimates in the first column of Table 4.2 are based on linear models equivalent to those in Equations 4.23 to 4.24, with income measured in real dollars. The second column provides estimates from models of essentially the same form, but with income transformed using the natural logarithm, denoted by $\ln(M)$, to correct this variable's positive skew and accommodate potential nonlinearities. The estimates in the third column extend those from the second–they are based on models that additionally include all two-way interactions between $C^{\perp}$ and D, $C^{\perp}$ and $\ln(M)$, and $L^{\perp}$ and $\ln(M)$. Links to the code and data used for this analysis can be found in the table footnote.

The RWR estimates vary considerably across different model specifications. In particular, estimates from models without confounder-exposure or confounder-mediator interactions do not point toward a strong mediating role for income. In contrast, estimates from models that include these interactions suggest that income could be an important mediator in the causal process linking college attendance to depression.

For example, consider the estimates from our most flexible specification, reported in the third column of the table. The estimate for the $OE(1,0)$ suggests that college attendance reduces CES-D scores by .112 standard deviations overall. The estimate for the $IDE(1,0)$ indicates that attending college would reduce CES-D scores by .041 standard deviations, on average, even if individuals received an income randomly selected from the distribution that would have emerged had everyone not completed college. In addition, the estimate for the $IIE(1,0)$ suggests that if individuals received an income randomly selected from the distribution that would have emerged had everyone attended college, rather than from the distribution that would have arisen had they not attended college, their level of depression would decline by .071 standard deviations.

Taken together, these estimates imply that income may partially mediate the overall effect of college attendance on depression. However, the estimate for the $CDE(1,0,50\text{K})$ indicates that CES-D scores would still decline by .084 standard deviations as a result of attending college, even if everyone were to receive a "middle class" income of $50,000. The difference between the $CDE(1,0,50\text{K})$ and $IDE(1,0)$ suggests that education and income might interact to produce their effects on depression. To further evaluate this possibility, we could compare a range of controlled direct effects across different values for the mediator, or we could more closely examine the coefficient on the exposure-mediator interaction in our outcome regression.

In conclusion, interventional and controlled direct effects can be estimated using the method of RWR when exposure-induced confounders are present. This approach yields consistent estimates if the assumptions required for identifying the effects of interest are met and if models for the mediator, outcome, and exposure-induced confounders are correctly specified. RWR can be implemented with any GLM for the exposure-induced confounders, but it requires a linear model for the outcome. It also typically utilizes a linear model for the mediator, although this is not strictly necessary. In the next section, we introduce another versatile approach for estimating interventional and controlled direct effects that can accommodate a wide range of linear and nonlinear models for the mediator, outcome, and an exposure-induced confounder.

4.6.2 Estimation via Simulation

In this section, we explain how to estimate interventional and controlled direct effects using a simulation approach compatible with any generalized linear model (GLM) for the mediator, outcome, and exposure-induced confounders (Zhou and Wodtke 2024). This approach builds upon the simulation estimator presented in Section 3.5.2 of the previous chapter, adapting it to applications with exposure-induced confounders. After first outlining a basic implementation of the simulation approach with a single exposure-induced confounder, we then explain how it can be extended to handle multiple exposure-induced confounders as well.

The *simulation estimator* is implemented through a series of steps. First, GLMs are fit to the sample data for the mediator, the outcome, and an exposure-induced confounder. Next, the models for the mediator and the exposure-induced confounder are used to simulate values for these variables under different exposures. Then, the outcome model is used to simulate values for the outcome under different exposures, while taking into account the simulated values for the mediator and exposure-induced confounder. Finally, the simulated outcomes are used to construct estimates for the interventional effects of interest.

Specifically, when estimating the $IDE\left(d, d^*\right)$, $IIE\left(d, d^*\right)$, and $OE\left(d, d^*\right)$, this procedure is implemented as follows:

1. **Fit models for the mediator, outcome, and exposure-induced confounder.** That is, first fit GLMs for the mediator and exposure-induced confounder, given the baseline confounders and the exposure. These models can be represented by $g\left(M|C, D\right)$ and $q\left(L|C, D\right)$, respectively. Then, fit another GLM for the outcome, given the baseline confounders, exposure, exposure-induced confounder, and mediator, which can be represented by $h\left(Y|C, D, L, M\right)$. Let $\hat{g}\left(M|C, D\right)$, $\hat{q}\left(L|C, D\right)$, and $\hat{h}\left(Y|C, D, L, M\right)$ denote these models with their parameters estimated by maximum likelihood.

2. **Simulate values for the exposure-induced confounder.** For each individual in the sample, simulate J copies of $L\left(d^*\right)$ and $L\left(d\right)$ from $\hat{q}\left(L|C, d^*\right)$ and $\hat{q}\left(L|C, d\right)$, respectively. Let $\tilde{L}_j\left(d^*\right)$ and $\tilde{L}_j\left(d\right)$ denote these values for each simulation $j = 1, 2, \ldots, J$.

3. **Simulate values for the mediator.** For each individual in the sample, simulate J copies of $\mathcal{M}\left(d^*|C\right)$ and $\mathcal{M}\left(d|C\right)$ from $\hat{g}\left(M|C, d^*\right)$ and $\hat{g}\left(M|C, d\right)$, respectively. Let $\tilde{\mathcal{M}}_j\left(d^*|C\right)$ and $\tilde{\mathcal{M}}_j\left(d|C\right)$ denote these values for each simulation $j = 1, 2, \ldots, J$.

4. **Simulate potential outcomes.** For every sample member and each set of simulated values for the mediator and exposure-induced confounder, simulate one copy of $Y\left(d, \mathcal{M}\left(d|C\right)\right)$ from $\hat{h}\left(Y|C, d, \tilde{L}_j\left(d\right), \tilde{\mathcal{M}}_j\left(d|C\right)\right)$, one copy of $Y\left(d, \mathcal{M}\left(d^*|C\right)\right)$ from $\hat{h}\left(Y|C, d, \tilde{L}_j\left(d\right), \tilde{\mathcal{M}}_j\left(d^*|C\right)\right)$, and one

copy of $Y\left(d^*, \mathcal{M}\left(d^*|C\right)\right)$ from $\hat{h}\left(Y|C, d^*, \tilde{L}_j\left(d^*\right), \tilde{\mathcal{M}}_j\left(d^*|C\right)\right)$. Let $\tilde{Y}_j\left(d, \mathcal{M}\left(d|C\right)\right)$, $\tilde{Y}_j\left(d, \mathcal{M}\left(d^*|C\right)\right)$, and $\tilde{Y}_j\left(d^*, \mathcal{M}\left(d^*|C\right)\right)$ denote these simulated outcomes for each simulation $j = 1, 2, \ldots, J$.

5. **Compute effect estimates.** Estimators for the interventional direct, interventional indirect, and overall effects are given by the following functions of the simulated outcomes:

$$\begin{aligned}
\widehat{IDE}\left(d, d^*\right)^{sim} &= \frac{1}{nJ}\sum\sum_{j}\left(\tilde{Y}_j\left(d, \mathcal{M}\left(d^*|C\right)\right) - \tilde{Y}_j\left(d^*, \mathcal{M}\left(d^*|C\right)\right)\right) \\
\widehat{IIE}\left(d, d^*\right)^{sim} &= \frac{1}{nJ}\sum\sum_{j}\left(\tilde{Y}_j\left(d, \mathcal{M}\left(d|C\right)\right) - \tilde{Y}_j\left(d, \mathcal{M}\left(d^*|C\right)\right)\right) \\
\widehat{OE}\left(d, d^*\right)^{sim} &= \frac{1}{nJ}\sum\sum_{j}\left(\tilde{Y}_j\left(d, \mathcal{M}\left(d|C\right)\right) - \tilde{Y}_j\left(d^*, \mathcal{M}\left(d^*|C\right)\right)\right), \qquad (4.27)
\end{aligned}$$

where the inner sum is taken over the J simulations, and the outer sum is taken over the n sample members. The "sim" superscript in these expressions denotes that they are simulation estimators.

In the initial step of this estimation procedure, we first fit a GLM for the mediator, using the baseline confounders and the exposure as predictors. Next, we fit a GLM for the exposure-induced confounder, also using the baseline confounders and the exposure as predictors. Then, we fit another GLM for the outcome, now with the baseline confounders, exposure, exposure-induced confounder, and mediator all included as predictors. The simulation approach can be implemented with any combination of GLMs for these variables. The key to its successful implementation lies in the correct specification of these models, as the consistency of the resulting effect estimates depends on it.

In step 2, we use the fitted model for the exposure-induced confounder to simulate potential values for this variable under different exposures. For each sample member, we replace their exposure with the value d^* but leave their observed values for the baseline confounders unchanged. With the predictors set at these levels, we then we select J Monte Carlo samples of the exposure-induced confounder from our fitted model. These simulated values are denoted by $\tilde{L}_j\left(d^*\right)$ for $j = 1, 2, \ldots, J$, and they serve as estimates for the potential value of the exposure-induced confounder under exposure d^*. This process is then repeated by replacing each sample member's exposure with the value d, again leaving their baseline confounders unchanged, and selecting another set of J Monte Carlo samples from our fitted model. These are denoted by $\tilde{L}_j\left(d\right)$ for $j = 1, 2, \ldots, J$, and represent estimates for the potential values of the exposure-induced confounder under exposure d.

Similarly, in step 3, we use our fitted model for the mediator to simulate random draws from its distribution under different exposures. For each individual in the sample, we replace their exposure with the value d^*, while keeping their observed values on the baseline confounders unchanged. Using our fitted model, we then select J Monte Carlo samples with the predictors set at these levels. The simulated values are denoted by $\tilde{\mathcal{M}}_j\left(d^*|C\right)$ for $j = 1, 2, \ldots, J$, and they approximate random draws of the mediator from its distribution under exposure d^*. We then repeat this process by replacing each sample member's exposure with the value d, again leaving the baseline confounders at their observed levels, and selecting another set of J Monte Carlo samples from our fitted model. These simulated values are denoted by $\tilde{\mathcal{M}}_j\left(d|C\right)$ for $j = 1, 2, \ldots, J$. They approximate random draws of the mediator from its distribution under the other exposure d.

In step 4, we use our fitted model for the outcome to simulate values under different levels of the exposure, mediator, and exposure-induced confounder. For each sample member, we begin by replacing their exposure with the value d, their mediator with one of its simulated values $\tilde{\mathcal{M}}_j\left(d|C\right)$, and their exposure-induced confounder with one of its simulated values $\tilde{L}_j(d)$, while keeping the baseline confounders at their observed levels. With the predictors set at these values, we then select one Monte Carlo sample of the outcome from our fitted model. This process is repeated for each combination of simulated values for the mediator and exposure-induced confounder. The resulting set of simulated outcomes is denoted by $\tilde{Y}_j(d, \mathcal{M}(d|C))$ for $j = 1, 2, \ldots, J$. These values serve as estimates of each individual's potential outcome under exposure d, with the mediator selected at random from its distribution under that same exposure.

Next, step 4 continues by keeping each sample member's exposure at the value d and leaving the baseline confounders at their observed levels. At this juncture, we again replace each individual's exposure-induced confounder with a corresponding simulated value $\tilde{L}_j(d)$, but we now reset their mediator to a simulated value drawn from its distribution under the alternative exposure, given by $\tilde{\mathcal{M}}_j(d^*|C)$. With the predictors set at these values, we then select one Monte Carlo sample of the outcome from our fitted model. Repeating this procedure for each combination of $\tilde{L}_j(d)$ and $\tilde{\mathcal{M}}_j(d^*|C)$ generates a set of simulated values for the outcome, denoted by $\tilde{Y}_j(d, \mathcal{M}(d^*|C))$ for $j = 1, 2, \ldots, J$. These values represent estimates of a sample member's potential outcome under exposure d, with the mediator randomly selected from its distribution under the alternative exposure d^*.

Step 4 then proceeds by resetting each sample member's exposure at the value d^*, and by replacing their exposure-induced confounder and mediator with $\tilde{L}_j(d^*)$ and $\tilde{\mathcal{M}}_j(d^*|C)$, respectively. With the predictors set at these levels, we select another Monte Carlo sample from our fitted model for the outcome, iterating this procedure across each combination of simulated values for the mediator and exposure-induced confounder. The resulting set of samples, denoted by $\tilde{Y}_j(d^*, \mathcal{M}(d^*|C))$ for $j = 1, 2, \ldots, J$, represent estimates of a sample member's potential outcome under exposure d^*, with the mediator randomly selected from its distribution under that same exposure. At the conclusion of step 4, we are left with J simulated values for each of the randomized potential outcomes that define the interventional effects of interest.

Lastly, in step 5, we compare these simulated outcomes to construct our desired effect estimates. This involves calculating differences between specific outcomes and then averaging these differences together across both simulations and sample members. For example, to compute an estimate of the $IDE(d, d^*)$, we evaluate the difference between $\tilde{Y}_j(d, \mathcal{M}(d^*|C))$ and $\tilde{Y}_j(d^*, \mathcal{M}(d^*|C))$ across each simulation and sample member. Averaging all of these differences together, then, yields a point estimate for the interventional direct effect. The procedures to estimate the $IIE(d, d^*)$ and $OE(d, d^*)$ are essentially identical but involve different contrasts among the simulated outcomes.

To appreciate the logic of the simulation approach, it is helpful to consider its relationship to the identification formulas described earlier. For example, elaborating on Equation 4.16 from Section 4.4, the identification formula for the marginal expected value of the randomized potential outcome $Y(d, \mathcal{M}(d^*|C))$ can be expressed as follows:

$$\begin{aligned}\mathbb{E}\left[Y\left(d, \mathcal{M}\left(d^*|C\right)\right)\right] &= \sum_{l,m,c} \mathbb{E}\left[Y|c,d,l,m\right] P\left(l|c,d\right) P\left(m|c,d^*\right) P\left(c\right) \\ &= \sum_{c} \Big(\sum_{y,l,m} yP\left(y|c,d,l,m\right) P\left(l|c,d\right) P\left(m|c,d^*\right) \Big) P\left(c\right). \qquad (4.28)\end{aligned}$$

For continuous data, the probability-weighted sums are replaced by density-weighted integrals, but with either continuous or discrete data, the identification formula depends on the conditional distributions of the mediator, the exposure-induced confounder, and the outcome. These distributions are denoted by $P(m|c, d^*)$, $P(l|c, d)$, and $P(y|c, d, l, m)$, respectively, in the expression above. The simulation approach models these distributions parametrically, generates Monte Carlo samples from them, and then averages these samples together, as outlined previously. By averaging across simulations, this procedure approximates the sum over y, l, and m in Equation 4.28. Averaging the resulting quantities again across sample members approximates the sum over c, using the empirical distribution of the confounders as an estimate for $P(c)$. Monte Carlo sampling from distribution models for M, L, and Y offers a relatively simple way to approximate the sums (or integrals) in the identification formula, especially when evaluating them directly is complicated by the use of nonlinear or otherwise complex models for the mediator, exposure-induced confounder, and/or the outcome.

It is important to underscore another key aspect of the simulation approach outlined here. Whenever simulating potential outcomes under a particular level of the exposure, we always align the exposure-induced confounder with one of its simulated values corresponding to that same level of exposure. This alignment ensures that our estimates are appropriately averaged over the conditional distribution of the exposure-induced confounder, following the nonparametric identification formulas in Equations 4.13 to 4.15. By extension, aligning simulated values for the confounder and outcome in this way safeguards against biases that could otherwise result from over-control or endogenous selection.

A similar procedure can be used to estimate controlled direct effects. Specifically, when estimating the $CDE\left(d, d^{*}, m\right)$ in the presence of an exposure-induced confounder, the simulation approach is implemented as follows:

1. **Fit models for the outcome and exposure-induced confounder.** That is, fit a GLM for the exposure-induced confounder, given the baseline confounders and the exposure, denoted by $q\left(L|C, D\right)$. Then, fit another GLM for the outcome, given the baseline confounders, exposure, exposure-induced confounder, and mediator, represented by $h\left(Y|C, D, L, M\right)$. The fitted models, with their parameters estimated by maximum likelihood, are denoted as $\hat{q}\left(L|C, D\right)$ and $\hat{h}\left(Y|C, D, L, M\right)$.

2. **Simulate values for the exposure-induced confounder.** For each individual in the sample, create J simulations of $L\left(d^{*}\right)$ from $\hat{q}\left(L|C, d^{*}\right)$, and another J simulations of $L\left(d\right)$ from $\hat{q}\left(L|C, d\right)$. Let $\tilde{L}_{j}\left(d^{*}\right)$ and $\tilde{L}_{j}\left(d\right)$ denote these values for each simulation $j = 1, 2, \ldots, J$.

3. **Simulate potential outcomes.** For every individual and each simulated value of the exposure-induced confounder, simulate one copy of $Y\left(d, m\right)$ from $\hat{h}\left(Y|C, d, \tilde{L}_{j}\left(d\right), m\right)$ and one copy of $Y\left(d^{*}, m\right)$ from $\hat{h}\left(Y|C, d^{*}, \tilde{L}_{j}\left(d^{*}\right), m\right)$. Let $\tilde{Y}_{j}\left(d, m\right)$ and $\tilde{Y}_{j}\left(d^{*}, m\right)$ denote these outcomes for each simulation $j = 1, 2, \ldots, J$.

4. **Compute the effect estimate.** An estimator for the controlled direct effect is given by the following function of the simulated outcomes:

$$\widehat{CDE}\left(d, d^{*}, m\right)^{sim} = \frac{1}{nJ} \sum \sum_{j} \left(\tilde{Y}_{j}\left(d, m\right) - \tilde{Y}_{j}\left(d^{*}, m\right)\right), \tag{4.29}$$

   where the inner sum is taken over the J simulations, and the outer sum is taken over the n sample members. As before, the "sim" superscript indicates that this expression is a simulation-based estimator.

In the procedure to estimate controlled direct effects, we begin by fitting two GLMs: one for the exposure-induced confounder with the exposure and baseline confounders as predictors, and another for the outcome using the baseline confounders, exposure, exposure-induced confounder, and mediator as predictors. Next, we simulate values for the exposure-induced confounder, exactly as we did for the interventional effects above. We then use our fitted model for the outcome to simulate values under different exposures but the same level of the mediator. Specifically, for each sample member, we replace their exposure with d^*, their mediator with m, and their exposure-induced confounder with one of its simulated values $\tilde{L}_j\left(d^*\right)$, leaving the baseline confounders unchanged. We then select one Monte Carlo sample from our fitted model, iterating across simulated values for the exposure-induced confounder to yield a set of simulated outcomes, $\tilde{Y}_j\left(d^*, m\right)$ for $j = 1, 2, \ldots, J$. Following this, we replace each sample member's exposure and exposure-induced confounder with d and $\tilde{L}_j\left(d\right)$, respectively, keeping the baseline confounders unchanged and mediator at level m. With the predictors now set at these values, we obtain another set of simulated outcomes, denoted $\tilde{Y}_j\left(d, m\right)$ for $j = 1, 2, \ldots, J$. Finally, we compute differences between the simulated outcomes, averaging them over simulations and sample members to estimate the controlled direct effect.

In Section 3.5.2 of Chapter 3, where we analyzed mediation in the absence of exposure-induced confounding, readers may recall that estimating the controlled direct effect required only a single model to compute predicted values for the outcome. There was no need to simulate outcomes using Monte Carlo samples from a series of distribution models. This simplification was possible because, with only baseline confounding, the identification formula for the controlled direct effect does not involve averaging over the distribution of any intermediate variables affected by the exposure. However, when exposure-induced confounding is present, the identification formula for the controlled direct effect does require averaging over such variables, as shown in Equation 4.9. The simulation approach outlined here addresses this requirement by modeling both $P\left(l|c, d\right)$ and $P\left(y|c, d, l, m\right)$, generating Monte Carlo samples from these distribution models, and then averaging these samples together. As with natural effects, Monte Carlo sampling is used to approximate the calculations specified in the identification formula.

To provide a more concrete illustration of the simulation approach, consider our empirical example based on the NLSY. In this analysis, the aim is to estimate the effects of college attendance on depression, as mediated by income. Additionally, the analysis adjusts for a set of baseline confounders and an exposure-induced confounder–unemployment status. How might we implement the simulation estimator with these data?

We first need to specify and fit GLMs for the exposure-induced confounder (unemployment status), mediator (income), and outcome (CES-D scores). Because unemployment is binary, we model its distribution using a logistic regression. Income, being continuous and positively skewed, is modeled using a log-normal linear regression, while CES-D scores are modeled with a normal linear regression. These models can be formally expressed as follows:

$$q\left(L|c, d\right) = Bern\left(p = logit^{-1}\left(\lambda_0 + \lambda_1^T c + \lambda_2 d\right)\right) \tag{4.30}$$

$$g\left(\ln\left(M\right)|c, d\right) = Norm\left(\mu_{\ln(M)} = \beta_0 + \beta_1^T c + \beta_2 d, \sigma^2_{\ln(M)}\right) \tag{4.31}$$

$$h\left(Y|c, d, l, m\right) = Norm\left(\mu_Y = \gamma_0 + \gamma_1^T c + \gamma_2 d + \ln\left(m\right)\left(\gamma_3 + \gamma_4 d\right) + \gamma_5 l, \sigma_Y^2\right), \tag{4.32}$$

where $\ln\left(M\right)$ denotes the natural logarithm of income, $logit^{-1}\left(\cdot\right) = \frac{exp(\cdot)}{1+exp(\cdot)}$ is the inverse logit function, $Bern\left(p\right)$ represents the Bernoulli probability distribution, and $Norm\left(\mu, \sigma^2\right)$ refers to the normal distribution with mean μ and variance σ^2.

After fitting these models by maximum likelihood, we proceed to simulate values for the mediator and exposure-induced confounder. We first replace each sample member's exposure with $d^* = 0$, and then select $J = 2000$ Monte Carlo samples of $\tilde{L}_j\left(0\right)$ from a Bernoulli distribution with $\hat{p} = logit^{-1}\left(\hat{\lambda}_0 + \hat{\lambda}_1^T C\right)$, along

Table 4.3: Interventional Effects of College Attendance on CES-D Scores as Estimated from the NLSY using the Simulation Approach.

Estimand	Point Estimates	
	Logit Model for L; Linear Model for $\ln(M)$; Linear Model for Y with $D \times \ln(M)$ Interaction	Logit Model for L; Linear Model for $\ln(M)$; Linear Model for Y; All Two-way Interactions
$OE(1,0)$	$-.077$	$-.117$
$IDE(1,0)$	$-.040$	$-.041$
$IIE(1,0)$	$-.037$	$-.075$
$CDE(1,0,50\text{K})$	$-.028$	$-.085$

Note: Estimates are expressed in standard deviation units. D denotes college attendance, L denotes unemployment status, $\ln(M)$ denotes the natural log of income, and Y denotes CES-D scores. Results are based on $J = 2000$ simulations. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch4/table_4-3`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

with samples of $\ln\left(\tilde{\mathcal{M}}_j(0|C)\right)$ from a normal distribution with mean $\hat{\mu}_{\ln(M)} = \hat{\beta}_0 + \hat{\beta}_1^T C$ and variance $\hat{\sigma}^2_{\ln(M)} = MSE_{\ln(M)}$. Next, we replace each sample member's exposure with $d = 1$, and select another $J = 2000$ Monte Carlo samples of $\tilde{L}_j(1)$ from a Bernoulli distribution with $\hat{p} = logit^{-1}\left(\hat{\lambda}_0 + \hat{\lambda}_1^T C + \hat{\lambda}_2\right)$, along with samples of $\ln\left(\tilde{\mathcal{M}}_j(1|C)\right)$ from a normal distribution with mean $\hat{\mu}_{\ln(M)} = \hat{\beta}_0 + \hat{\beta}_1^T C + \hat{\beta}_2$ and the same variance as before. In these expressions, the "hats" denote maximum likelihood estimates, and $MSE_{\ln(M)}$ represents the mean squared error from our model for $\ln(M)$.

With simulated values for the mediator and exposure-induced confounder, we now use these to simulate values for the outcome. For each sample member and each simulation j, we select values for $\tilde{Y}_j(1, \mathcal{M}(1|C))$, $\tilde{Y}_j(1, \mathcal{M}(0|C))$, and $\tilde{Y}_j(0, \mathcal{M}(0|C))$ from a series of normal distributions with the same variance but different means. Specifically, these distributions have means given by $\hat{\mu}_Y = \hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_2 + \ln\left(\tilde{\mathcal{M}}_j(1|C)\right)(\hat{\gamma}_3 + \hat{\gamma}_4) + \gamma_5 \tilde{L}_j(1)$, $\hat{\mu}_Y = \hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_2 + \ln\left(\tilde{\mathcal{M}}_j(0|C)\right)(\hat{\gamma}_3 + \hat{\gamma}_4) + \gamma_5 \tilde{L}_j(1)$, and $\hat{\mu}_Y = \hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_3 \ln\left(\tilde{\mathcal{M}}_j(0|C)\right) + \gamma_5 \tilde{L}_j(0)$, respectively, all with variance equal to $\hat{\sigma}^2_Y = MSE_Y$, the mean squared error from our outcome model. Finally, to construct estimates for the interventional effects of interest, we substitute these simulated outcomes into the expressions for $\widehat{IDE}(1,0)^{sim}$, $\widehat{IIE}(1,0)^{sim}$, and $\widehat{OE}(1,0)^{sim}$, as in Equation 4.27 above.

To estimate controlled direct effects, we simulate a distinct set of potential outcomes. For each sample member and each of the j simulated values for the exposure-induced confounder, we select Monte Carlo samples of $\tilde{Y}_j(1, 50\text{K})$ and $\tilde{Y}_j(0, 50\text{K})$ from two different normal distributions. The first distribution has a mean of $\hat{\mu}_Y = \hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_2 + \ln(50\text{K})(\hat{\gamma}_3 + \hat{\gamma}_4) + \gamma_5 \tilde{L}_j(1)$, while the second has a mean of $\hat{\mu}_Y = \hat{\gamma}_0 + \hat{\gamma}_1^T C + \hat{\gamma}_3 \ln(50\text{K}) + \gamma_5 \tilde{L}_j(0)$. Both distributions share the same variance, $\hat{\sigma}^2_Y = MSE_Y$. By substituting these simulated outcomes into the expression for $\widehat{CDE}(1,0,50\text{K})^{sim}$, following Equation 4.27, we obtain an estimate for the controlled direct effect of interest.

Results from this analysis are reported in the first column of Table 4.3. The estimates in the second column follow a similar procedure, with one key difference: the models for income, depression, and unemployment additionally include two-way interactions among the predictors. This modification allows the effects of both the exposure and mediator to vary across levels of the confounders. Links to the code and data used for this analysis are provided in the table footnote.

Estimates from the simulation approach are broadly in line with those based on regression-with-residuals (RWR), which we reported previously in Table 4.2. They too highlight a notable degree of variability across different model specifications. In particular, the estimates derived from models without interactions between the confounders and exposure, or between the confounders and mediator, do not point toward a very strong mediating role for income. Conversely, when these interactions are included, our effect estimates do suggest that income acts as an important mediator connecting college attendance to lower levels of depression at midlife. For example, in the second column of the table, our estimates for the $IDE\,(1,0)$, $IIE\,(1,0)$, and $OE\,(1,0)$ are $-.041$, $-.076$, and $-.117$ standard deviations, respectively. The estimated indirect effect is not only substantively large but also constitutes a sizable component of the overall effect.

To summarize, interventional and controlled direct effects can be estimated by fitting GLMs to sample data and then by using these models to simulate potential outcomes. This approach will yield accurate estimates provided the necessary assumptions for identifying these effects are met, and the models assumed for the mediator, outcome, and exposure-induced confounder are correctly specified. If these conditions hold, the simulation estimators outlined previously will converge to their target estimands as both the sample size n and number of Monte Carlo simulations J increase without limit. Importantly, the analyst has full control over the number of Monte Carlo simulations, and the only consideration when increasing J is the additional computational resources required to manage the extra simulations. In practice, we suggest $10^3 \leq J \leq 10^4$, as this should be sufficient to minimize simulation error in most applications.

The simulation estimators discussed in this section offer considerable flexibility, as they can be implemented with any GLM for the mediator, the outcome, and an exposure-induced confounder. Within the family of GLMs, there exists an extensive array of models capable of accommodating many different types of variables, their associated probability distributions, and the diverse functional relationships among them. As such, the simulation approach is a highly versatile and adaptable method of parametric estimation for interventional and controlled direct effects. The method's versatility and adaptability make it particularly well-suited for applications with binary, ordinal, or count variables, whether they serve as the outcome, mediator, or an exposure-induced confounder.

The simulation approach can also be adapted for applications with multiple exposure-induced confounders. In such cases, it is necessary to model the *joint* distribution of these variables, conditional on the baseline confounders and exposure, in order to properly simulate their potential values. A joint probability distribution describes all the possible values and their associated probabilities for two or more random variables. To model the joint distribution of multiple exposure-induced confounders, we can use the product rule of joint probability. This rule allows us to decompose the joint distribution into a series of conditional distributions. Each conditional distribution can then be modeled separately using an appropriate GLM.

Specifically, to accommodate a multivariate set of exposure-induced confounders, we modify the steps used to implement the simulation estimator as follows. In the first step, all the exposure-induced confounders are included as predictors in the GLM for the outcome. This model can be then denoted as $h\,(Y|C, D, \mathbf{L}, M)$, where $\mathbf{L} = (L_1, L_2, \ldots, L_K)$ represents a vector of K exposure-induced confounders. In addition, we now fit a series of GLMs for each exposure-induced confounder. These models can be denoted by $\hat{q_1}\,(L_1|C, D)\,, \hat{q_2}\,(L_2|C, D, L_1)\,, \ldots, \hat{q_K}\,(L_K|C, D, \mathbf{L}_{K-1})$, where $\mathbf{L}_{K-1} = (L_1, L_2, \ldots, L_{K-1})$. With this approach, the variables in $\mathbf{L}$ are ordered arbitrarily and then a GLM is fit for each, sequentially, using the baseline confounders, exposure, and all preceding exposure-induced confounders as predictors.

In the second step, we implement a sequential simulation procedure to generate potential values for all the exposure-induced confounders. Initially, we simulate J copies of $L_1\,(d^*)$ and $L_1\,(d)$ from $\hat{q_1}\,(L_1|C, d^*)$

and $\hat{q}(L_1|C,d)$, respectively. The simulated values can be denoted as $\tilde{L}_{1j}(d^*)$ and $\tilde{L}_{1j}(d)$ for each simulation $j=1,2,\ldots,J$. Next, we simulate J copies of $L_2(d^*)$ and $L_2(d)$ from $\hat{q_2}\left(L_2|C,d^*,\tilde{L}_{1j}(d^*)\right)$ and $\hat{q}\left(L_2|C,d,\tilde{L}_{1j}(d)\right)$, which can be denoted by $\tilde{L}_{2j}(d^*)$ and $\tilde{L}_{2j}(d)$ for $j=1,2,\ldots,J$. This procedure continues for each exposure-induced confounder in sequence, using the simulated values of preceding confounders to generate simulations for subsequent ones. Ultimately, this yields complete vectors of simulated confounders under exposures d and d^*. These vectors can be denoted as $\tilde{\mathbf{L}}_j(d)=\left(\tilde{L}_{1j}(d),\tilde{L}_{2j}(d),\ldots,\tilde{L}_{Kj}(d)\right)$ and $\tilde{\mathbf{L}}_j(d^*)=\left(\tilde{L}_{1j}(d^*),\tilde{L}_{2j}(d^*),\ldots,\tilde{L}_{Kj}(d^*)\right)$ for each simulation $j=1,2,\ldots,J$.

Finally, the vectors of simulated confounders are carried forward to the step in which potential outcomes are simulated from the GLM for Y. For example, to simulate copies of $Y(d,\mathcal{M}(d|C))$, we replace each sample member's observed exposure with the value d, their observed mediator with a simulated value $\tilde{\mathcal{M}}_j(d|C)$, and their vector of exposure-induced confounders with the set of simulated values $\tilde{\mathbf{L}}_j(d)$, and then we select Monte Carlo samples from $\hat{h}\left(Y|C,d,\tilde{\mathbf{L}}_j(d),\tilde{\mathcal{M}}_j(d|C)\right)$. Simulated values for the other potential outcomes are generated analogously. Aside from these modifications, the simulation approach proceeds exactly as outlined previously for both interventional and controlled direct effects.

Although the simulation approach is easily adapted for applications with multiple exposure-induced confounders, modeling their joint distribution can still be challenging in practice. These difficulties stem from the need to correctly specify a series of GLMs for each confounder. Misspecification in any one of these models can introduce bias, which may limit the utility of the simulation approach in settings with many exposure-induced confounders. In these settings, estimation via simulation can become unwieldy and susceptible to model misspecification as the number of confounders grows large. It is therefore best suited to applications with a relatively small number of exposure-induced confounders. In applications with many such variables, RWR is generally the preferred approach, as it can incorporate them more easily.

4.6.3 Estimation with Inverse Probability Weights

In this section, we explain how to estimate interventional effects using inverse probability weights (VanderWeele et al. 2014). This approach is an extension of the weighting estimators discussed in the previous chapter and is designed for applications with a single exposure-induced confounder. Although inverse probability weighting can accommodate multiple exposure-induced confounders in theory, its implementation becomes prohibitively complex with the inclusion of more than a few such variables.

In the presence of an exposure-induced confounder, *weighting estimators* for the $IDE(d,d^*)$, $IIE(d,d^*)$, and $OE(d,d^*)$ can be constructed through the following series of steps. First, generalized linear models (GLMs) for the exposure, exposure-induced confounder, and mediator are fit to the sample data. Next, the fitted models are used to estimate a set of probabilities for the exposure, confounder, and mediator. Based on these probabilities, weights are then constructed to transform the empirical distribution of the sample data so that it emulates the intended results of certain hypothetical experiments. Finally, the interventional effects of interest are estimated by comparing the mean of the outcome across differently weighted samples.

More specifically, the procedure for estimating interventional effects via inverse probability weighting is implemented as follows:

1. **Fit a model for the exposure and predict probabilities.** That is, fit a GLM for the exposure, conditional on the baseline confounders, denoted by $f(D|C)$. For each sample member, use the maximum likelihood fit for this model, $\hat{f}(D|C)$, to predict the probability of exposure to d and the probability of exposure to d^*. Let $\hat{P}(d|C)$ and $\hat{P}(d^*|C)$ denote each of these predicted probabilities.

2. **Fit a model for the exposure-induced confounder and predict probabilities.** Fit another GLM for the exposure-induced confounder, conditional on the exposure and baseline confounders, denoted by $q\left(L|C,D\right)$. For each sample member, use the maximum likelihood fit for this model, $\hat{q}\left(L|C,D\right)$, to predict the probability of each level l on the exposure-induced confounder, conditional on their observed values for the baseline confounders and exposure to d. Then, use $\hat{q}\left(L|C,D\right)$ to predict the probability of each level l on the exposure-induced confounder, conditional on a sample member's observed values for the baseline confounders and exposure to d^*. Let $\hat{P}\left(l|C,d\right)$ and $\hat{P}\left(l|C,d^*\right)$ denote these predicted probabilities.

3. **Fit a model for the mediator and predict probabilities.** Fit another GLM for the mediator, conditional on the baseline confounders, exposure, and exposure-induced confounder, denoted by $g\left(M|C,D,L\right)$. Use the maximum likelihood fit for this model, $\hat{g}\left(M|C,D,L\right)$, to predict the probability of a sample member's observed level for the mediator, conditional on their observed values for the baseline confounders, level l of the exposure-induced confounder, and level d of the exposure. In addition, use $\hat{g}\left(M|C,D,L\right)$ to predict the probability of a sample member's observed level for the mediator, conditional on their observed values for the baseline confounders, level l of the exposure-induced confounder, and level d^* of the exposure. These predicted probabilities are computed across each level l of the exposure-induced confounder, and they are denoted by $\hat{P}\left(M|C,d^*,l\right)$ and $\hat{P}\left(M|C,d,l\right)$.

4. **Construct inverse probability weights.** Among sample members exposed to d^*, compute a set of inverse probability weights given by

$$\hat{w}_1 = \frac{\sum_l \hat{P}\left(M|C,d^*,l\right)\hat{P}\left(l|C,d^*\right)}{\hat{P}\left(d^*|C\right)\hat{P}\left(M|C,d^*,L\right)}.$$

For those exposed to d, compute two additional sets of inverse probability weights given by

$$\hat{w}_2 = \frac{\sum_l \hat{P}\left(M|C,d,l\right)\hat{P}\left(l|C,d\right)}{\hat{P}\left(d|C\right)\hat{P}\left(M|C,d,L\right)} \text{ and } \hat{w}_3 = \frac{\sum_l \hat{P}\left(M|C,d^*,l\right)\hat{P}\left(l|C,d^*\right)}{\hat{P}\left(d|C\right)\hat{P}\left(M|C,d,L\right)}.$$

5. **Compute effect estimates.** Estimators for the interventional direct, indirect, and overall effects are given by the following contrasts between weighted means of the outcome:

$$\begin{aligned}
\widehat{IDE}\left(d,d^*\right)^{ipw} &= \frac{\sum I\left(D=d\right)\hat{w}_3 Y}{\sum I\left(D=d\right)\hat{w}_3} - \frac{\sum I\left(D=d^*\right)\hat{w}_1 Y}{\sum I\left(D=d^*\right)\hat{w}_1} \\
\widehat{IIE}\left(d,d^*\right)^{ipw} &= \frac{\sum I\left(D=d\right)\hat{w}_2 Y}{\sum I\left(D=d\right)\hat{w}_2} - \frac{\sum I\left(D=d\right)\hat{w}_3 Y}{\sum I\left(D=d\right)\hat{w}_3} \\
\widehat{OE}\left(d,d^*\right)^{ipw} &= \frac{\sum I\left(D=d\right)\hat{w}_2 Y}{\sum I\left(D=d\right)\hat{w}_2} - \frac{\sum I\left(D=d^*\right)\hat{w}_1 Y}{\sum I\left(D=d^*\right)\hat{w}_1},
\end{aligned} \tag{4.33}$$

where the "ipw" superscript denotes that they are based on inverse probability weights. In these expressions, $I\left(\cdot\right)$ is an indicator function equal to 1 when its argument is true, and 0 otherwise. By extension, $\sum I(D=d^*)\hat{w}_1 Y / \sum I(D=d^*)\hat{w}_1$ is a weighted mean of the outcome Y among sample members for whom $D = d^*$, with weights given by $\hat{w}_1$. The expressions $\sum I(D=d)\hat{w}_2 Y / \sum I(D=d)\hat{w}_2$ and $\sum I(D=d)\hat{w}_3 Y / \sum I(D=d)\hat{w}_3$ are defined analogously.

In step 1 of this estimation procedure, we first fit a GLM for the exposure, using only the baseline confounders as predictors. We then use this model to predict the probability that each sample member is exposed to d, given their observed values on the baseline confounders. We also use this model to predict the probability that each sample member is exposed to d^*, conditional on their baseline confounders. These predicted probabilities are denoted by $\hat{P}(d|C)$ and $\hat{P}(d^*|C)$, respectively.

In step 2, we fit a GLM for the exposure-induced confounder. This model includes both the baseline confounders and the exposure as predictors. Using this fitted model, we predict the probability of each level for the exposure-induced confounder, given a sample member's observed values for the baseline confounders and exposure to d. We also predict the probability of each level for the exposure-induced confounder, now given a sample member's observed values for the baseline confounders and exposure to d^*. These predicted probabilities are denoted by $\hat{P}(l|C,d)$ and $\hat{P}(l|C,d^*)$, respectively, for each level l of the confounder.

In step 3, we a fit a GLM for the mediator, using the baseline confounders, exposure, and exposure-induced confounder as predictors. Based on this model, we then predict the probability of a sample member's observed level for the mediator, given their observed values for the baseline confounders, level l of the exposure-induced confounder, and level d of the exposure. In addition, we also predict the probability of a sample member's observed mediator, given their baseline confounders, level l of the exposure-induced confounder, and level d^* of the exposure. We compute these probabilities across each level l of the exposure-induced confounder, and denote them by $\hat{P}(M|C,d^*,l)$ and $\hat{P}(M|C,d,l)$.

In step 4, we construct inverse probability weights. The first set of weights is given by

$$\hat{w}_1 = \frac{\sum_l \hat{P}(M|C,d^*,l)\,\hat{P}(l|C,d^*)}{\hat{P}(d^*|C)\,\hat{P}(M|C,d^*,L)}.$$

Among sample members for whom $D = d^*$, weighting by $\hat{w}_1$ transforms the empirical distribution of the data so that exposure to d^* appears to have been randomly assigned. It also transforms the data so that the mediator for each individual appears to have been randomly drawn from its distribution under exposure d^*.

Similarly, the second set of weights is given by

$$\hat{w}_2 = \frac{\sum_l \hat{P}(M|C,d,l)\,\hat{P}(l|C,d)}{\hat{P}(d|C)\,\hat{P}(M|C,d,L)}.$$

For those in the sample with $D = d$, weighting by $\hat{w}_2$ transforms the empirical distribution of the data so that exposure to d appears to have occurred at random. Additionally, it also transforms the data so that the mediator for each individual appears to have been chosen randomly from its distribution under exposure d.

The third set of weights is given by

$$\hat{w}_3 = \frac{\sum_l \hat{P}(M|C,d^*,l)\,\hat{P}(l|C,d^*)}{\hat{P}(d|C)\,\hat{P}(M|C,d,L)}.$$

Among sample members exposed to d, weighting by $\hat{w}_3$ transforms the empirical distribution of the data so that this exposure seems to have been randomly assigned. It also transforms the data so that the mediator for each individual appears to have been randomly selected from its distribution under the alternative exposure d^*.

In step 5, we compute the weighted mean of the outcome within each of these subsamples, and then take differences between them to estimate the interventional effects of interest. Specifically, the weighted mean of the outcome in the subsample exposed to d^*, given by $\sum I(D=d^*)\hat{w}_1 Y / \sum I(D=d^*)\hat{w}_1$, serves as an estimate for $\mathbb{E}[Y(d^*,\mathcal{M}(d^*|C))]$. This is because weighting by $\hat{w}_1$ creates the appearance that exposure to d^* was randomly assigned and that the mediator was randomly selected from its distribution under the same exposure.

Similarly, weighting the subsample exposed to d by $\hat{w}_2$ creates the appearance that this exposure was randomly assigned. It also creates the appearance that the mediator was chosen at random from its distribution under exposure d. As a result, the weighted mean of the outcome, given by $\sum I(D=d)\hat{w}_2 Y / \sum I(D=d)\hat{w}_2$, here serves as an estimate for $\mathbb{E}\left[Y\left(d, \mathcal{M}\left(d|C\right)\right)\right]$.

Finally, weighting the same subsample now by $\hat{w}_3$ again emulates random assignment to exposure d. Using this new weight, however, additionally creates the appearance that the mediator was randomly selected from its distribution under the alternative exposure d^*. As a result, the weighted mean of the outcome, denoted by $\sum I(D=d)\hat{w}_3 Y / \sum I(D=d)\hat{w}_3$, in this case provides an estimate for $\mathbb{E}\left[Y\left(d, \mathcal{M}\left(d^*|C\right)\right)\right]$. Comparing these weighted means, as outlined in Equation 4.33, yields estimates for the interventional effects of interest.

A similar procedure can be used to estimate controlled direct effects in the presence of exposure-induced confounding (VanderWeele 2009b). When focusing on the $CDE\left(d, d^*, m\right)$, we only fit GLMs for the exposure and mediator, but not for any exposure-induced confounders. Next, we use the fitted models to predict a set of exposure and mediator probabilities. With these probabilities, we then construct weights that transform the sample data to emulate an experiment in which the exposure and mediator have been jointly randomized. Finally, we estimate controlled direct effects by comparing outcome means in the weighted sample.

Specifically, to estimate the $CDE\left(d, d^*, m\right)$ when exposure-induced confounders are present, inverse probability weighting is implemented as follows:

1. **Fit models for the exposure and mediator.** That is, fit a GLM for the exposure, given the baseline confounders, denoted by $f\left(D|C\right)$. Fit another GLM for the mediator, given the baseline confounders, the exposure, and exposure-induced confounders, denoted by $g\left(M|C, D, L\right)$. Let $\hat{f}\left(D|C\right)$ and $\hat{g}\left(M|C, D, L\right)$ represent these models with their parameters estimated by maximum likelihood.

2. **Compute predicted probabilities.** Use $\hat{f}\left(D|C\right)$ to predict the probability of each sample member's observed exposure, given their baseline confounders. In addition, use $\hat{g}\left(M|C, D, L\right)$ to predict the probability of a sample member's observed level for the mediator, given their baseline confounders, exposure, and exposure-induced confounders. Let $\hat{P}\left(D|C\right)$ and $\hat{P}\left(M|C, D, L\right)$ denote each of these predicted probabilities.

3. **Construct inverse probability weights.** Calculate an inverse probability weight for each sample member, given by
$$\hat{w}_4 = \frac{1}{\hat{P}\left(M|C, D, L\right)\hat{P}\left(D|C\right)}.$$

4. **Compute effect estimates.** An estimator for the controlled direct effect is given by the following contrast between weighted means of the observed outcome:
$$\widehat{CDE}\left(d, d^*, m\right)^{ipwl} = \frac{\sum I\left(D=d, M=m\right)\hat{w}_4 Y}{\sum I\left(D=d, M=m\right)\hat{w}_4} - \frac{\sum I\left(D=d^*, M=m\right)\hat{w}_4 Y}{\sum I\left(D=d^*, M=m\right)\hat{w}_4}. \tag{4.34}$$

   In this expression, $\sum I(D=d, M=m)\hat{w}_4 Y / \sum I(D=d, M=m)\hat{w}_4$ is a weighted mean of the outcome Y among sample members for whom $D=d$ and $M=m$, with weights equal to $\hat{w}_4 = 1/\hat{P}(m|C,d,L)\hat{P}(d|C)$. Similarly, $\sum I(D=d^*, M=m)\hat{w}_4 Y / \sum I(D=d^*, M=m)\hat{w}_4$ is a weighted mean of the outcome among sample members with $D=d^*$ and $M=m$, where the weights are equal to $\hat{w}_4 = 1/\hat{P}(m|C,d^*,L)\hat{P}(d^*|C)$.

This procedure closely parallels the approach outlined in Section 3.5.3 from the previous chapter, which focused on estimating controlled direct effects without exposure-induced confounding. The weighting estimator outlined here differs only in that the exposure-induced confounders are included as additional predictors in the model for the mediator.

The weights used to estimate controlled direct effects are given by $\hat{w}_4 = {}^{1}\!/_{\hat{P}(M|C,D,L)\hat{P}(D|C)}$. They are the product of an inverse probability for the mediator, ${}^{1}\!/_{\hat{P}(M|C,D,L)}$, and another for the exposure, ${}^{1}\!/_{\hat{P}(D|C)}$. Weighting the sample by $\hat{w}_4$ balances the distribution of the baseline confounders across levels of both the exposure and mediator. Additionally, it also balances the distribution of the exposure-induced confounders across levels of the mediator, within levels of the exposure. In other words, $\hat{w}_4$ transforms the sample so that it appears as though the data came from an experiment where the exposure and mediator were jointly randomized.

To elaborate further, applying these weights to the subsample of individuals for whom $D = d$ and $M = m$ makes it appear as though level d of the exposure and level m of the mediator were randomly assigned. As a result, the weighted mean of the outcome in this subsample, given by ${}^{\sum I(D=d,M=m)\hat{w}_4 Y}\!/_{\sum I(D=d,M=m)\hat{w}_4}$, serves as an estimate for $\mathbb{E}\left[Y\left(d,m\right)\right]$. Similarly, when the weights are applied to the subsample for whom $D = d^*$ and $M = m$, it emulates random assignment for these levels of the exposure and mediator. Thus, the weighted mean of the outcome in this subsample, represented by ${}^{\sum I(D=d^*,M=m)\hat{w}_4 Y}\!/_{\sum I(D=d^*,M=m)\hat{w}_4}$, serves as an estimate for $\mathbb{E}\left[Y\left(d^*,m\right)\right]$. The difference between these two weighted means, as detailed in Equation 4.34, provides an estimate for the controlled direct effect.

Alternatively, if either the exposure D or mediator M have many values, we could also estimate the controlled direct effect using a parametric model for $\mathbb{E}\left[Y|d,m\right]$ together with the inverse probability weights. For example, we could fit a linear model for the outcome, such as $\mathbb{E}\left[Y|d,m\right] = \upsilon_0 + \upsilon_1 d + m\left(\upsilon_2 + \upsilon_3 d\right)$,using the method of weighted least squares and weights equal to $\hat{w}_4$. Weighted least squares is very similar to ordinary least squares except it involves selecting estimates for model parameters by finding values that minimize the *weighted* sum of squared prediction errors (c.f., Fox 2015). Because weighting the sample by $\hat{w}_4$ appropriately balances the distribution of the confounders across the exposure and mediator, adjustment for these variables in the outcome model is no longer necessary, and an estimator for the controlled direct effect could be obtained by calculating $\widehat{CDE}\left(d,d^*,m\right)^{ipwl} = \left(\hat{\upsilon}_1 + \hat{\upsilon}_3 m\right)\left(d - d^*\right)$ from the model's coefficients. Models of this form are known as *marginal structural models* (Robins et al., 2000; VanderWeele, 2009b).

Inverse probability weighting mitigates exposure-induced confounding without introducing bias from over-control or endogenous selection. This is because the weights only balance the exposure-induced confounders across levels of the mediator, not across levels of the exposure itself. Figure 4.8 presents a stylized graph illustrating the effect of re-weighting the observed data. Specifically, Panel A depicts the relationships among variables in the unweighted data, while Panel B highlights how these relationships are transformed after weighting by $\hat{w}_4$.

When the data are weighted, the figure illustrates that the exposure D is no longer related to the baseline confounders C, and that the mediator M is no longer related to either the baseline confounders C or the exposure-induced confounders L. Importantly, however, the relationship between the exposure D and the exposure-induced confounders L remains intact. Inverse probability weighting essentially neutralizes the causal paths emanating from C into D and M, and from L into M, while leaving the path from D into L unchanged. As a result, it can be used to adjust for both baseline and exposure-induced confounders, without incurring any bias due to over-control or endogenous selection.

Inverse probability weighting yields consistent estimates for interventional and controlled direct effects, provided that the assumptions required for identifying these effects are satisfied and the models used to construct the weights are correctly specified. If any identification assumptions are not met (e.g., due to un-

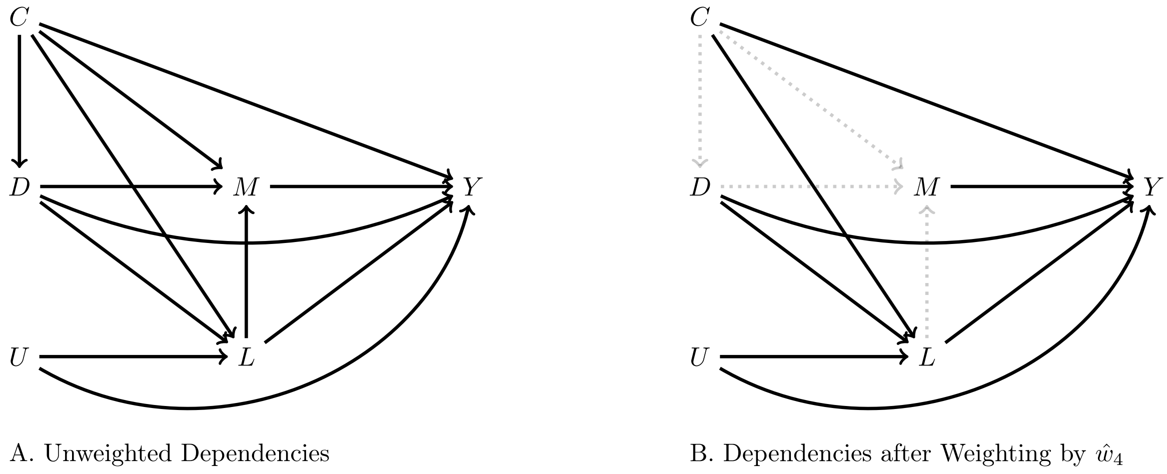

C
D
M
Y
U
L
A. Unweighted Dependencies
C
D
M
Y
U
L
B. Dependencies after Weighting by $\hat{w}_4$

Figure 4.8: Graphical Illustration of Adjustment for Confounding with Inverse Probability Weights.

Note: D denotes the exposure, M denotes a focal mediator of interest, L denotes an exposure-induced confounder, Y denotes the outcome, C denotes a set of baseline confounders, and U denotes an unobserved variable.

observed confounding), these estimates will suffer from bias. Moreover, even if all identification assumptions hold, model misspecification can still result in bias, as weights constructed from incorrect models may not transform the sample data as intended.

The weights defined previously involve inverse probabilities that can be highly variable and sometimes produce extreme values, leading to imprecise and unstable effect estimates. These difficulties can be partly mitigated by using stabilized versions of the weights, which are scaled down to have a mean of 1 and smaller variance.

A set of stabilized weights for estimating interventional effects are given by the following expressions:

$$\hat{sw}_1 = \hat{w}_1 \times \hat{P}\left(d^*\right), \ \hat{sw}_2 = \hat{w}_2 \times \hat{P}\left(d\right), \text{ and } \hat{sw}_3 = \hat{w}_3 \times \hat{P}\left(d\right), \tag{4.35}$$

where $\hat{P}\left(d^*\right)$ and $\hat{P}\left(d\right)$ are marginal probabilities of exposure to d^* and d, respectively. Similarly, a stabilized weight for estimating controlled direct effects can be expressed as follows:

$$\hat{sw}_4 = \hat{wt}_4 \times \hat{P}\left(M|D\right) \hat{P}\left(D\right), \tag{4.36}$$

where $\hat{P}\left(M|D\right)$ is the predicted probability of a sample member's observed value for the mediator, given only their exposure, and $\hat{P}\left(D\right)$ is the marginal probability of a sample member's observed exposure.

To construct effect estimates based on the stabilized weights, analysts need only compute $\{\hat{sw}_1, \hat{sw}_2, ..., \hat{sw}_4\}$ and then substitute these weights for $\{\hat{w}_1, \hat{w}_2, ..., \hat{w}_4\}$ in Equations 4.33 and 4.34 above. Substituting stabilized for unstabilized weights in the estimation procedures outlined previously will tend to yield effect estimates that are less variable and more precise.

Moreover, as with the weighting estimators introduced in Chapter 3, censoring the weights can also improve their performance, provided that it is not applied excessively. For example, researchers might consider censoring the weights at their 1st and 99th percentiles, which will usually decrease the variability of effect estimates without introducing appreciable bias.

Table 4.4: Interventional Effects of College Attendance on CES-D Scores as Estimated from the NLSY using Inverse Probability Weighting.

Estimand	Point Estimates	
	Logit Model for D; Logit Model for L; Linear Model for $\ln(M)$; No Interactions	Logit Model for D; Logit Model for L; Linear Model for $\ln(M)$; All Two-way Interactions
$OE(1,0)$	$-.177$	$-.178$
$IDE(1,0)$	$-.159$	$-.159$
$IIE(1,0)$	$-.018$	$-.019$
$CDE(1,0,50\text{K})$	$-.147$	$-.150$

Note: Estimates are expressed in standard deviation units. D denotes college attendance, L denotes unemployment status, and $\ln(M)$ denotes the natural log of income. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch4/table_4-4`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

Although inverse probability weighting is compatible with a wide range of GLMs, it is best suited for applications where the exposure and mediator are discrete and have a limited number of values. For variables with many values or that are continuous, the method tends to be unreliable, as seemingly minor errors in model specification can lead to large distortions in the weights and, consequently, badly biased effect estimates. In general, inverse probability weighting performs optimally when the exposure and mediator are binary, ordinal, or polytomous.

Furthermore, when researchers are interested in estimating interventional effects, the method is also best suited for applications involving a single exposure-induced confounder with relatively few values. If multiple such confounders are present, or even a single one with many values, it becomes increasingly difficult to accurately estimate the numerator of the weights, and small specification errors can again lead to large biases in effect estimates. Thus, when targeting the $IDE(d,d^*)$, $IIE(d,d^*)$, and $OE(d,d^*)$ in particular, inverse probability weighting should be used mainly for analyses with just one exposure-induced confounder that is also binary, ordinal, or polytomous. For other applications, researchers should consider alternative methods like regression-with-residuals (RWR) or the simulation approach.

Using the NLSY, we applied inverse probability weighting to estimate the interventional and controlled direct effects of college attendance on CES-D scores, with income as our focal mediator and unemployment status as a potential exposure-induced confounder. We used logit models for college attendance and for unemployment status, and we used a log-normal model for income. We fit these models by maximum likelihood, both with and without two-way interactions between the confounders and the exposure. We then used them to construct a set of stabilized weights, which we also censored at the 1st and 99th percentiles to further improve precision.

Results from this analysis are presented in Table 4.4, with replication files accessible via links in the table footnote. Compared to our earlier analyses based on RWR and the simulation approach, these estimates suggest larger direct and overall effects, but a smaller indirect effect. This pattern of results remains fairly stable, whether or not we include interactions in the models used to construct the weights. Thus, estimates based on inverse probability weighting provide relatively little evidence that income mediates the influence of college attendance on later life depression.

However, the divergent results given by RWR, simulation, and weighting suggests that this conclusion is

Table 4.5: Bootstrap Inferential Statistics for the Interventional Effects of College Attendance on CES-D Scores Computed from the NLSY.

Estimands	RWR Estimates		Simulation Estimates		IPW Estimates	
	P-value	95% CI	P-value	95% CI	P-value	95% CI
$OE(1,0)$	.019	$[-.205, -.019]$	.018	$[-.207, -.022]$	.005	$[-.289, -.051]$
$IDE(1,0)$	.455	$[-.144, .060]$	.457	$[-.144, .060]$	.023	$[-.275, -.021]$
$IIE(1,0)$	$< .001$	$[-.113, -.036]$	$< .001$	$[-.122, -.037]$	.232	$[-.060, .011]$
$CDE(1,0,50\text{K})$	.077	$[-.176, .009]$	.077	$[-.176, .009]$	.008	$[-.256, -.042]$

Note: Bootstrap p-values and confidence intervals are based on $B = 2000$ replications. P-values are from tests of the null hypothesis that the focal estimand is equal to zero. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch3/table_4-5`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

sensitive to the distinct modeling assumptions required of each approach. In this situation, it may be prudent to place greater trust in the method or methods that best complement the specific features of our analysis with the NLSY, while minimizing the impact of inherent limitations. Given that inverse probability weighting can perform poorly with continuous mediators like income, its results should be interpreted cautiously, and we might consider prioritizing the more consistent findings from RWR and the simulation approach, which are better equipped to handle mediators with many values.

4.7 Statistical Inference

To compute inferential statistics for interventional and controlled direct effects when exposure-induced confounding is present, we employ the same methods outlined in Chapter 3. Specifically, we rely on the nonparametric bootstrap to form confidence intervals and conduct hypothesis tests for the effects in question.

Briefly, the nonparametric bootstrap is a technique for approximating the sampling distribution of an estimator. This approximation is achieved by repeatedly drawing samples from the available data with replacement and then computing separate estimates from each sample. Under fairly general conditions, the estimates obtained from these multiple samples closely correspond with the true but unknown sampling distribution (Davison and Hinkley 1997; Efron and Tibshirani 1994). Then, armed with an approximate sampling distribution, we can construct confidence intervals and evaluate null hypotheses. For a more comprehensive discussion of the nonparametric bootstrap and related concepts, refer to Section 3.6 in the preceding chapter.

Table 4.5 presents inferential statistics from our analysis of the NLSY, calculated using the nonparametric bootstrap. Specifically, it presents 95% confidence intervals for the interventional and controlled direct effects of college attendance on depression, with income as our focal mediator and unemployment as an exposure-induced confounder.

We constructed these confidence intervals by drawing $B = 2000$ bootstrap samples, computing effect estimates for each, and then finding the 2.5th and 97.5th percentiles of the resulting estimates. The intervals provide a range of values that contain the target estimand with 95% probability under repeated sampling, assuming no model misspecification and that the estimand is identifiable. In general, wider confidence intervals signal a higher degree of uncertainty in our results due to the variability inherent in random sampling.

Table 4.5 also includes p-values from tests of the null hypothesis that the effects of interest are equal to zero. These results quantify the likelihood of observing a point estimate as or even more extreme than what we actually obtained from our sample data, assuming the null hypothesis is true. Small p-values therefore signal that the sample data are inconsistent with a null hypothesis, while larger values indicate that there is little evidence against it. We calculated these p-values by inverting bootstrap confidence intervals. Specifically, we constructed a bootstrap distribution of $B = 2000$ estimates and then located the smallest value of $(100-\tau)/100$ where the hypothesized value of zero falls outside the corresponding τ-percent confidence interval.

The results in Table 4.5 suggest that income likely mediates the effect of college attendance on depression. For example, when testing the null hypothesis of no interventional indirect effect, both regression-with-residuals and the simulation approach yield p-values close to zero. This means that our sample data provide considerable evidence against the possibility that income plays no mediating role. Additionally, the confidence intervals suggest that the indirect effect is negative and moderate in size.

Similarly, null hypothesis tests for the overall and controlled direct effects also produce small p-values, while findings for the interventional direct effect are less consistent. However, the wide range of the confidence intervals for all these effects, regardless of the estimation method employed, points to a high degree of sampling variability.

4.8 Sensitivity Analysis

While the methods discussed in this chapter accommodate both baseline and exposure-induced confounders that are observed, they are still premised on the assumption that there is no unobserved confounding of the exposure-outcome, exposure-mediator, and mediator-outcome relationships. If unobserved confounders do exist, then the estimators discussed previously would be biased, and they would not converge to their target estimands with increases in sample size. Consequently, our statistical inferences would be invalid, and our conclusions about causal mediation could be mistaken, regardless of how much data we collect.

In this section, we introduce a series of bias formulas for interventional direct and indirect effects in the presence of unobserved confounding. These formulas allow us to evaluate the sensitivity of estimates to hypothetical patterns of unobserved confounding and to explore how the resulting bias may alter our conclusions about causal mediation. As in the previous chapter, we first present bias formulas in a general form that imposes few constraints on the pattern of confounding. Despite their generality, these formulas involve a large number of parameters, making them difficult to use in practical applications. Thus, after introducing the bias formulas in their most general form, we then consider particular patterns of unobserved confounding, which, when combined with some additional simplifying assumptions, allow the bias formulas to be simplified as straightforward functions of a few basic sensitivity parameters. This simplified approach is easier to use in practice because it requires specifying only a handful of parameters, although it hinges on stronger assumptions about the nature of unobserved confounding.

4.8.1 Nonparametric Bias Formulas

Consider first the scenario in which an unobserved variable, denoted by U, affects the exposure D, the mediator M, the outcome Y, and the exposure-induced confounders L. Because the unobserved variable confounds the exposure-outcome, mediator-outcome, and exposure-mediator relationships, all of the conditional independence assumptions required to identify interventional direct and indirect effects are violated, and by extension, all of the estimators we have considered thus far would be biased and inconsistent. In this situation, the bias afflicting an estimator for the controlled direct effect is given by

$$\begin{aligned}\text{Bias}\left(\widehat{CDE}\left(d,d^{*},m\right)\right) =& \sum_{l,u,c}\mathbb{E}[Y|c,d,l,m,u]P(l|c,d)P(c)\left(P(u|c,d,l,m)-\frac{P(u|c,d,l)P(u|c)}{P(u|c,d)}\right)\\ &-\sum_{l,u,c}\mathbb{E}[Y|c,d^{*},l,m,u]P(l|c,d^{*})P(c)\\ &\times\left(P(u|c,d^{*},l,m)-\frac{P(u|c,d^{*},l)P(u|c)}{P(u|c,d^{*})}\right).\end{aligned}\tag{4.37}$$

The bias afflicting an estimator for the interventional direct effect is given by

$$\begin{aligned}\text{Bias}\left(\widehat{IDE}\left(d,d^{*}\right)\right) =& \sum_{l,m,u,c}\mathbb{E}[Y|c,d,l,m,u]P(l|c,d)P(m|c,d^{*})P(c)\\ &\times\left(P(u|c,d,l,m)-\frac{P(u|c,d,l)P(u|c,d^{*},m)P(u|c)}{P(u|c,d)P(u|c,d^{*})}\right)\\ &-\sum_{l,m,u,c}\mathbb{E}[Y|c,d^{*},l,m,u]P(l|c,d^{*})P(m|c,d^{*})P(c)\\ &\times\left(P(u|c,d^{*},l,m)-\frac{P(u|c,d^{*},l)P(u|c,d^{*},m)P(u|c)}{P(u|c,d^{*})P(u|c,d^{*})}\right).\end{aligned}\tag{4.38}$$

And the bias afflicting an estimator for the interventional indirect effect is given by

$$\begin{aligned}\text{Bias}\left(\widehat{IIE}\left(d,d^{*}\right)\right) =& \sum_{l,m,u,c}\mathbb{E}[Y|c,d,l,m,u]P(l|c,d)P(m|c,d)P(c)\\ &\times\left(P(u|c,d,l,m)-\frac{P(u|c,d,l)P(u|c,d,m)P(u|c)}{P(u|c,d)P(u|c,d)}\right)\\ &-\sum_{l,m,u,c}\mathbb{E}[Y|c,d,l,m,u]P(l|c,d)P(m|c,d^{*})P(c)\\ &\times\left(P(u|c,d,l,m)-\frac{P(u|c,d,l)P(u|c,d^{*},m)P(u|c)}{P(u|c,d)P(u|c,d^{*})}\right).\end{aligned}\tag{4.39}$$

The bias afflicting an estimator for the overall effect is the sum of equations 4.38 and 4.39, $\text{Bias}\left(\widehat{OE}\left(d,d^{*}\right)\right) = \text{Bias}\left(\widehat{IDE}\left(d,d^{*}\right)\right)+\text{Bias}\left(\widehat{IIE}\left(d,d^{*}\right)\right)$.

The expressions presented above are highly complex. In general, they suggest that the biases affecting estimators for interventional effects depend on how the outcome differs across levels of the unobserved confounder given the exposure, mediator, exposure-induced confounders, and baseline confounders. They also suggest that the biases depend on how the unobserved confounder differs across levels of the exposure, mediator, and exposure-induced confounders. Specifying plausible values for all the different quantities that compose these bias terms is exceptionally challenging. Thus, it is usually necessary to consider more specific forms of confounding in order to facilitate the practical application of these bias expressions in a sensitivity analysis.

4.8.2 Bias from Exposure-outcome Confounding

Now consider a scenario in which an unobserved variable U confounds only the exposure-outcome relationship, but not the exposure-mediator or mediator-outcome relationships. This pattern of unobserved confounding may arise in observational studies where an unobserved variable is suspected to affect both the exposure and the outcome, but not the mediator. For example, when analyzing the role of income in mediating the effect of college attendance on mental health, it may be reasonable to assume that certain unobserved family characteristics, such as parental mental health, affect an individual's level of education and their own mental health later in life, but not their family income directly.

In addition, suppose that U is binary, that the difference in the mean of the outcome across levels of U does not depend on the exposure, mediator, or any of the confounders, and that the difference in the prevalence of U across levels of the exposure does not depend on the baseline confounders. Under these assumptions, the biases in estimators for the controlled and interventional direct effects are equal and can be expressed as follows:

$$\text{Bias}\left(\widehat{IDE}\left(d,d^*\right)\right) = \text{Bias}\left(\widehat{CDE}\left(d,d^*,m\right)\right) = \delta_{UY|C,D,L,M} \times \delta_{DU|C}, \tag{4.40}$$

where

$$\begin{aligned} \delta_{DU|C} &= P\left(U=1|c,d\right) - P\left(U=1|c,d^*\right) \\ \delta_{UY|C,D,L,M} &= \mathbb{E}\left[Y|c,d,l,m,U=1\right] - \mathbb{E}\left[Y|c,d,l,m,U=0\right]. \end{aligned}$$

The first sensitivity parameter in this expression, $\delta_{UY|C,D,L,M}$, represents the difference in the mean of the outcome between individuals with $U=1$ and those with $U=0$, conditional on the exposure, mediator, and all confounders. The second sensitivity parameter in this expression, $\delta_{DU|C}$, represents the difference in the prevalence of the unobserved confounder U comparing level d versus d^* of the exposure, conditional on the baseline confounders. Thus, under several simplifying assumptions, the bias in estimators for both interventional and controlled direct effects is equal to a "partial effect" of the unobserved confounder on the outcome ($\delta_{UY|C,D,L,M}$) multiplied by a "partial effect" of the exposure on the unobserved confounder ($\delta_{DU|C}$).

Under the same set of assumptions, estimators of the interventional indirect effect will be unbiased. This is because the interventional indirect effect captures a causal chain composed of the paths $D \rightarrow M$ and $M \rightarrow Y$, neither of which is subject to unobserved confounding in this case. Because the interventional indirect effect is unbiased in this situation, estimators for the overall effect will be subject to the same bias as estimators of the interventional direct effect. In sum, exposure-outcome confounding will distort our estimates of the controlled direct effect, interventional direct effect, and overall effect, but it is less likely to induce bias in estimates of the interventional indirect effect.

4.8.3 Bias from Mediator-outcome Confounding

Consider next a scenario in which an unobserved variable U confounds only the mediator-outcome relationship, but not the exposure-mediator or exposure-outcome relationship. This pattern of unobserved confounding may arise in an experimental study where only the exposure is randomly assigned, which would ensure that there is no unobserved confounding of the exposure-outcome or exposure-mediator relationships. However, random assignment of the exposure alone would not obviate the problem of mediator-outcome confounding by unobserved factors.

Suppose that U is binary and that the difference in the mean of the outcome across levels of U does not depend on the exposure, mediator, or any of the confounders, as before. In addition, suppose further that the difference in the prevalence of U across levels of the exposure does not depend on the baseline confounders or the mediator, after it has been averaged over the conditional distribution of the exposure-induced confounders–that is, assume that $\sum_l P\left(U=1|c,d,l,m\right)P\left(l|c,d\right)-\sum_l P\left(U=1|c,d^*,l,m\right)P\left(l|c,d^*\right)$ is a constant. Under these assumptions, the general bias formulas outlined previously can be simplified as follows:

$$\begin{aligned} \text{Bias}\left(\widehat{IDE}\left(d,d^*\right)\right) &= \text{Bias}\left(\widehat{CDE}\left(d,d^*,m\right)\right)=\delta_{UY|C,D,L,M}\times\delta_{DU|C,M} \\ \text{Bias}\left(\widehat{IIE}\left(d,d^*\right)\right) &= \delta_{UY|C,D,L,M}\times\delta_{DMU|C} \\ \text{Bias}\left(\widehat{OE}\left(d,d^*\right)\right) &= \left(\delta_{DU|C,M}+\delta_{DMU|C}\right)\delta_{UY|C,D,L,M}, \end{aligned} \tag{4.41}$$

where $\delta_{UY|C,D,L,M}$ is defined as before and

$$\begin{aligned} \delta_{DU|C,M} &= \sum_l P\left(U=1|c,d,l,m\right)P\left(l|c,d\right)-\sum_l P\left(U=1|c,d^*,l,m\right)P\left(l|c,d^*\right) \\ \delta_{DMU|C} &= \sum_{m,l,c} P\left(U=1|c,d,l,m\right)P\left(l|c,d\right)\left(P\left(m|c,d\right)-P\left(m|c,d^*\right)\right)P\left(c\right). \end{aligned}$$

The biases in estimators for the controlled and interventional direct effects are identical and equal to the product of two sensitivity parameters, $\delta_{UY|C,D,L,M}$ and $\delta_{DU|C,M}$. The first of these sensitivity parameters, $\delta_{UY|C,D,L,M}$, represents the difference in the mean of the outcome between individuals with $U=1$ and individuals with $U=0$, conditional on the exposure, mediator, and all confounders. The second sensitivity parameter, $\delta_{DU|C,M}$, is a measure of the association between the exposure and the unobserved confounder, conditional on the mediator and baseline confounders. If we were to treat the unobserved confounder U as a response variable in a regression, it would roughly correspond to an associational "direct effect" of D on U, adjusting for the mediator and baseline confounders. Thus, under several simplifying assumptions, the bias in estimators for both interventional and controlled direct effects is equal to a "partial effect" of the unobserved confounder on the outcome ($\delta_{UY|C,D,L,M}$) multiplied by an associational "direct effect" of the exposure on the unobserved confounder that adjusts for the mediator ($\delta_{DU|C,M}$).

Under the same set of assumptions, the bias in estimators for the interventional indirect effect is also equal to the product of two sensitivity parameters, $\delta_{UY|C,D,L,M}$ and $\delta_{DMU|C}$. The first of these sensitivity parameters is the same as that for the controlled and interventional direct effects, capturing the "partial effect" of the unobserved confounder on the outcome, conditional on the exposure, mediator, and all confounders. The second sensitivity parameter, $\delta_{DMU|C}$, is a measure of the association between the exposure and the unobserved confounder that is induced by the mediator. If we were to treat the unobserved confounder U as a response variable, it would roughly correspond to an associational "indirect effect" of D on U through the mediator M. Thus, the bias in estimators for the interventional indirect effect is equal to a "partial effect" of the unobserved confounder on the outcome ($\delta_{UY|C,D,L,M}$) multiplied by an associational "indirect effect" of the exposure on the unobserved confounder through the mediator ($\delta_{DMU|C}$).

Finally, the bias in an estimator for the overall effect is the sum of biases afflicting the interventional direct and indirect effects, which, in this case, is equal to $\left(\delta_{DU|C,M}+\delta_{DMU|C}\right)\delta_{UY|C,D,L,M}$. The sum of the two sensitivity parameters, $\delta_{DU|C,M}+\delta_{DMU|C}$, is a measure of the overall association between the exposure and the unobserved confounder, conditional on the baseline confounders.

4.8.4 Bias from Exposure-mediator Confounding

Finally, consider a scenario in which an unobserved variable U affects both the exposure and the mediator, but not the exposure-induced confounders or the outcome directly. In addition, suppose that the following simplifying assumptions hold: the unobserved confounder U is binary; the difference in the prevalence of U across levels of the exposure does not depend on the baseline confounders; the difference in the probability of the mediator across levels of the unobserved confounder does not depend on the exposure, conditional on the baseline confounders; and the controlled direct effect of the exposure on the outcome given the baseline confounders does not depend on the level of the mediator.

Under these assumptions, estimators for the controlled and interventional direct effects are unbiased. This is because the controlled direct effect captures the path $D \to Y$, which is not subject to unobserved confounding in this case. In fact, estimators for the controlled direct effect are unbiased even without the simplifying assumptions outlined previously. Without these assumptions, however, estimators for the interventional direct effect will generally be biased. This is because the interventional direct effect is a weighted average of controlled direct effects, with weights corresponding to the conditional distribution of $M(d^*)$, the potential value of the mediator under exposure to d^*, given the baseline confounders. Without adjusting for the unobserved confounder U, estimates for these weights based on the observed data will be incorrect, leading to biased estimates of the interventional direct effect. But if the controlled direct effect does not depend on the level of the mediator, the weights are irrelevant and estimators for the interventional direct effect will also be unbiased.

Under the same set of assumptions, estimators for the interventional indirect effect do suffer from bias. In this case, the bias in an estimator for the indirect effect is given by

$$\text{Bias}\left(\widehat{IIE}\left(d, d^*\right)\right) = \delta_{DU|C} \times \delta_{UMY|C}, \tag{4.42}$$

where $\delta_{DU|C}$ is defined as before and

$$\delta_{UMY|C} = \sum_{l,m,c} \mathbb{E}\left[Y|c, d, l, m\right] P\left(l|c, d\right) \left(P\left(m|c, d, U = 1\right) - P\left(m|c, d, U = 0\right)\right) P\left(c\right).$$

The first sensitivity parameter in this expression, $\delta_{DU|C}$, represents the difference in the prevalence of the unobserved confounder U across levels d versus d^* of the exposure, conditional on the baseline confounders. The second sensitivity parameter, $\delta_{UMY|C}$, is a measure of the association between the unobserved confounder and the outcome. It is similar to an “indirect effect” of the unobserved confounder on the outcome that operates through the mediator, conditional on the baseline confounders. Thus, the bias in estimators for the interventional indirect effect is equal to a “partial effect” of the exposure on the unobserved confounder ($\delta_{DU|C}$) multiplied by an associational “indirect effect” of the unobserved confounder on the outcome ($\delta_{UMY|C}$). Because the interventional direct effect is unbiased in this situation, estimators for the overall effect suffer the same bias as estimators of the interventional indirect effect.

Table 4.6: Simplified Bias Formulas for Interventional Effects.

Bias/estimator	Type of Confounding		
	$D \leftarrow U \rightarrow Y$	$M \leftarrow U \rightarrow Y$	$D \leftarrow U \rightarrow M$
Bias $\left(\widehat{CDE}\left(d, d^{*}, m\right)\right)$	$\delta_{UY\|C,D,L,M} \times \delta_{DU\|C}$	$\delta_{UY\|C,D,L,M} \times \delta_{DU\|C,M}$	0
Bias $\left(\widehat{IDE}\left(d, d^{*}\right)\right)$	$\delta_{UY\|C,D,L,M} \times \delta_{DU\|C}$	$\delta_{UY\|C,D,L,M} \times \delta_{DU\|C,M}$	0
Bias $\left(\widehat{IIE}\left(d, d^{*}\right)\right)$	0	$\delta_{UY\|C,D,L,M} \times \delta_{DMU\|C}$	$\delta_{DU\|C} \times \delta_{UMY\|C}$
Bias $\left(\widehat{OE}\left(d, d^{*}\right)\right)$	$\delta_{UY\|C,D,L,M} \times \delta_{DU\|C}$	$\left(\delta_{DU\|C,M} + \delta_{DMU\|C}\right) \delta_{UY\|C,D,L,M}$	$\delta_{DU\|C} \times \delta_{UMY\|C}$

Note: These bias formulas variously assume that U is binary; $P\left(U=1 \mid c, d\right) - P\left(U=1 \mid c, d^{*}\right)$ is constant across levels of C; $\mathbb{E}\left[Y \mid c, d, l, m, U=1\right] - \mathbb{E}\left[Y \mid c, d, l, m, U=0\right]$ is constant across levels of C, D, L, and M; $\sum_{l} P\left(U=1 \mid c, d, l, m\right) P\left(l \mid c, d\right) - \sum_{l} P\left(U=1 \mid c, d^{*}, l, m\right) P\left(l \mid c, d^{*}\right)$ is constant; $P\left(m \mid c, d, U=1\right) - P\left(m \mid c, d, U=0\right)$ is constant across levels of D; and for the bias in interventional direct effects under exposure-mediator confounding specifically, that the controlled direct effect does not depend on m within levels of the baseline confounders.

4.8.5 Bias-adjusted Effect Estimates

Table 4.6 summarizes the bias formulas for the interventional effects of interest under the simplifying assumptions outlined previously. With these formulas, a formal sensitivity analysis proceeds by reevaluating the focal effect estimates across different hypothetical patterns of unobserved confounding. To this end, we would begin by specifying the bias formulas with plausible values for their sensitivity parameters, and then we would construct a set of bias-adjusted effect estimates by subtracting the bias terms from their corresponding point estimates.

Specifically, a set of bias-adjusted estimates for the interventional effects of interest can be expressed as follows:

$$\begin{aligned}
\widehat{CDE}\left(d, d^{*}, m\right)^{adj} &= \widehat{CDE}\left(d, d^{*}\right) - \text{Bias}\left(\widehat{CDE}\left(d, d^{*}, m\right)\right) \\
\widehat{IDE}\left(d, d^{*}\right)^{adj} &= \widehat{IDE}\left(d, d^{*}\right) - \text{Bias}\left(\widehat{IDE}\left(d, d^{*}\right)\right) \\
\widehat{IIE}\left(d, d^{*}\right)^{adj} &= \widehat{IIE}\left(d, d^{*}\right) - \text{Bias}\left(\widehat{IIE}\left(d, d^{*}\right)\right) \\
\widehat{OE}\left(d, d^{*}\right)^{adj} &= \widehat{OE}\left(d, d^{*}\right) - \text{Bias}\left(\widehat{OE}\left(d, d^{*}\right)\right) \qquad (4.43)
\end{aligned}$$

where $\widehat{CDE}\left(d, d^{*}, m\right)$, $\widehat{IDE}\left(d, d^{*}\right)$, $\widehat{IIE}\left(d, d^{*}\right)$ and $\widehat{OE}\left(d, d^{*}\right)$ each denote an estimator from Sections 4.5 to 4.6, and the "adj" superscript indicates that they have been adjusted for bias due to an assumed pattern of unobserved confounding. We can assess the impact of unobserved confounding on our inferences by evaluating the bias-adjusted estimates across a range of plausible values for the sensitivity parameters. Confidence intervals for the bias-adjusted estimates can be constructed with the nonparametric bootstrap.

4.9 An Empirical Illustration: The Effect of Plow Use on Female Political Participation

In this section, we illustrate the methods outlined previously with a reanalysis of data from Alesina et al. (2013). This study sought to understand how pre-industrial farming practices may have shaped modern gender roles. The authors distinguished between labor-intensive forms of agriculture, using handheld tools like hoes, and capital-intensive methods that rely on plows, which require more upper-body strength to pull the plow or control the animal maneuvering it. Because of the physical strength required for plow cultivation, the authors hypothesized that men had a comparative advantage in farming, leading to a division of labor along gender lines in societies that practiced plow agriculture: men worked the fields, while women focused on domestic tasks. This division may have engendered cultural norms about the appropriate role of women in society, particularly the belief that they naturally belong in the home. The authors posited that these norms could endure long after a society has transitioned away from plow agriculture, thereby affecting female participation in contemporary politics (see also Boserup 1970 and Acharya et al. 2016).

To test this theory, Alesina et al. (2013) estimated the effect of historical plow use on the contemporary proportion of seats held by women in national parliament. Using country-level data and conventional regression analysis, they found that while the total effect of historical plow use was minimal, its direct impact became much larger and negative after adjusting for a country's per capita gross domestic product (GDP). They attribute these results to a combination of offsetting influences: a negative direct effect of plow use on female political participation, consistent with their hypothesis, and a positive indirect effect mediated by higher national incomes. As the authors explain (pg. 493-494):

> "Once we control for per capita income, the magnitude of the relationship between traditional plough use and female participation in politics roughly doubles and becomes statistically significant...even if traditional plough use has a negative impact on female participation in politics, countries with a tradition of plough use also tend to be richer, and therefore, all else equal, have more female participation in politics."

In our reanalysis, we further explore whether national income mediates the effect of historical plow use on female representation in government. In particular, we account for an exposure-induced confounder that could potentially influence these results–the extent to which a country's governance leans democratic or authoritarian.

The data for this analysis come from $n = 129$ countries around the world. The exposure D is a binary variable coded 1 for countries in which a majority of the population descends from communities that historically engaged in plow agriculture, and 0 for all other countries. The outcome Y represents the proportion of seats held by women in the single or lower chamber of a country's parliament as of the year 2000. The mediator of interest M is a country's per capita GDP, also measured in the year 2000. The vector of baseline confounders, denoted by C, is intended to control for historical differences between countries that adopted plow agriculture and those that did not. It includes measures capturing the availability of large domesticated animals, the agricultural suitability of a country's terrain, and whether the country has a tropical climate.

In addition, we also adjust for a potential exposure-induced confounder by incorporating a measure of authoritarian versus democratic governance, denoted by L. This measure is based on a country's "Polity Score" in the year 2000, which summarizes differences across several dimensions of political authority, including procedures for selecting executives, constraints on executive power, and the scope for political competition, among other factors (Marshall and Jaggers 2007). Using this Polity Score, we categorize each country into one of three governing styles: an "autocracy," a hybrid or mixed style termed "anocracy," or a "democracy."

Authoritarian versus democratic governance may serve as an exposure-induced confounder for several reasons. In general, a country's historical mode of production shape its political institutions, which in turn affect both representation in government and economic growth. For example, certain agricultural practices like plow cultivation may lead to more centralized forms of governance, which could then evolve into authoritarian rule. In authoritarian regimes, power is usually concentrated in a small, male-dominated group. These regimes may also foster a political culture that is less receptive to gender inclusivity and erect institutional barriers that limit roles for women. By contrast, in democratic societies, where political power is more diffuse and individual rights are better protected, inclusive and equitable representation tends to take higher priority, while the legal frameworks in these settings are more conducive to women's participation in politics. Thus, authoritarian versus democratic governance may be affected by pre-industrial modes of production and may in turn affect levels of both national income and female political representation.

In our reanalysis, we target several different effects of historical plow use. First, we focus on the interventional direct effect, which can be formally expressed as $IDE\left(1,0\right) = \mathbb{E}\left[Y\left(1,\mathcal{M}\left(0|C\right)\right) - Y\left(0,\mathcal{M}\left(0|C\right)\right)\right]$. This effect represents the expected difference in the proportion of parliamentary seats held by women if countries did, versus did not, historically utilize plow agriculture, while holding the distribution of national income constant at levels observed in the absence of plow cultivation. In essence, it isolates an effect of historical plow use on female participation in politics that is not due to differences in the distribution of GDP.

Next, we examine the interventional indirect effect, which can be formally expressed as $IIE\left(1,0\right) = \mathbb{E}\left[Y\left(1,\mathcal{M}\left(1|C\right)\right) - Y\left(1,\mathcal{M}\left(0|C\right)\right)\right]$. This effect represents the expected difference in female political participation if countries had a level of contemporary income drawn randomly from the distribution arising under plow cultivation rather than from the distribution under another type of pre-industrial agriculture. In other words, it captures an effect of historical plow use on women's participation in politics that operates through a mechanism involving national income.

We also consider the overall effect, defined as $OE\left(1,0\right) = \mathbb{E}\left[Y\left(1,\mathcal{M}\left(1|C\right)\right) - Y\left(0,\mathcal{M}\left(0|C\right)\right)\right]$. This effect captures the expected difference in female political participation due to the joint influence of historical plow cultivation and the change in the distribution of national income that came from the use of plows versus other forms of pre-industrial agriculture.

Lastly, we examine a controlled direct effect, formally defined as $CDE\left(1,0,1.8\text{K}\right) = \mathbb{E}\left[Y\left(1,1.8\text{K}\right) - Y\left(0,1.8\text{K}\right)\right]$. This effect captures the influence of historical plow use on the present-day share of parliamentary seats held by women if all countries had a per capita GDP around \$1800, which is roughly the sample average for the year 2000.

These effects can be identified if the following conditions are satisfied: (i) there is no unobserved confounding of the exposure-outcome and exposure-mediator relationships; (ii) there is no unobserved confounding of the mediator-outcome relationship; (iii) there is a positive probability of the exposure and mediator conditional on the confounders; and (iv) the observed values for the mediator and outcome are consistent with their potential values. None of these conditions are met by design in the data assembled by Alesina et al. (2013), and several are dubious. For example, there are many unmeasured factors that could confound the relationship between per capita GDP and female political participation, such as a country's level of investment in formal education, the vibrancy of its civil society, or the adequacy of its healthcare infrastructure. As a result, any causal inferences drawn from these data are provisional and must be interpreted very cautiously.

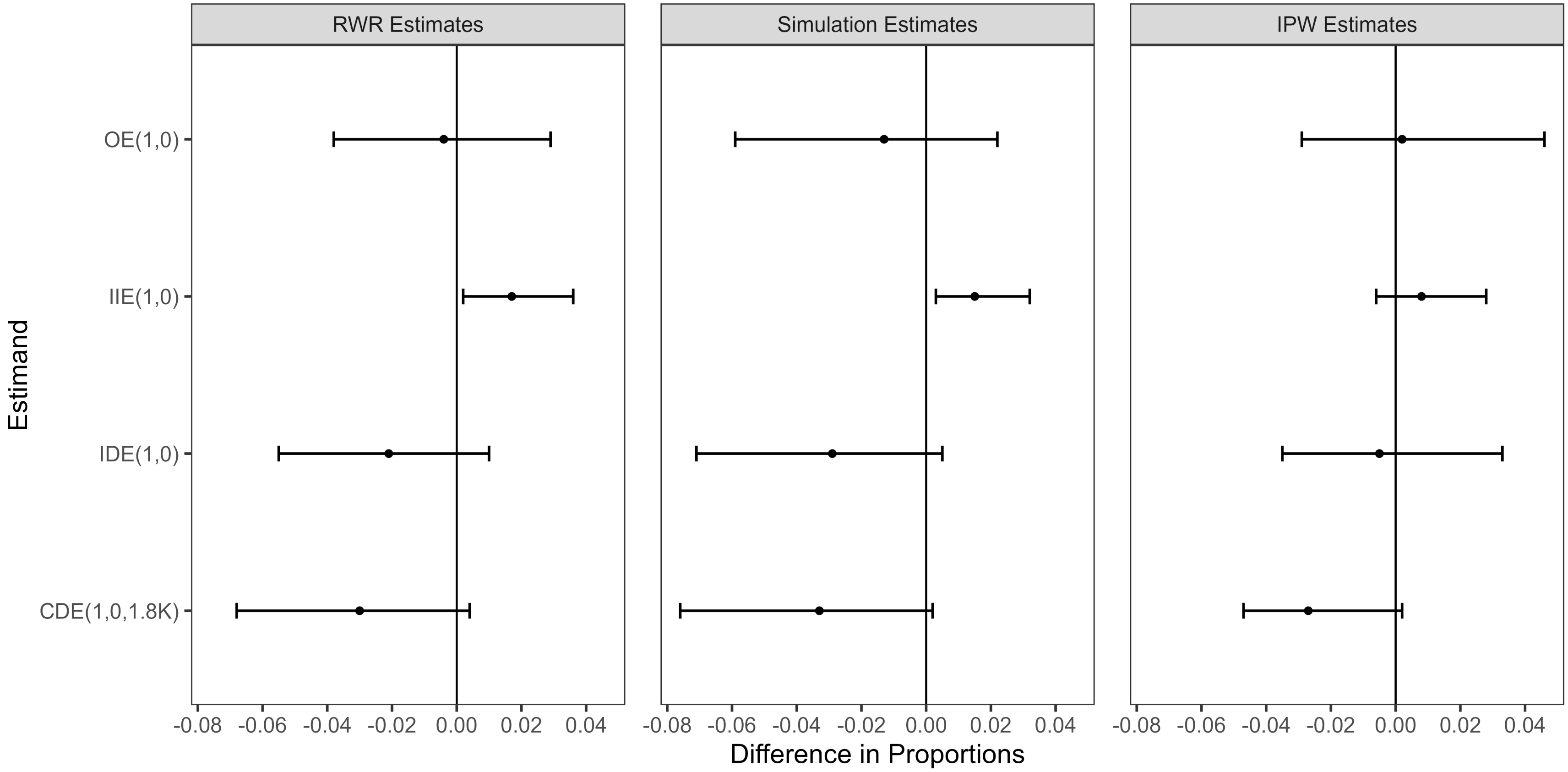


Figure 4.9: Interventional and Controlled Direct Effects of Historical Plow Use on Women's Contemporary Representation in Government.

Note: Estimates are expressed as differences in proportions. Bootstrap confidence intervals are based on $B = 2000$ replications. The regression-with-residual (RWR) estimates are based on additive linear models for $\mathbb{E}\left[M|C,D\right]$ and $\mathbb{E}\left[L|C,D\right]$, and a linear model for $\mathbb{E}\left[Y|C,D,L,M\right]$ with a with $D \times M$ interaction. The simulation estimates are based on an ordinal logit model for $q\left(L|C,D\right)$ (i.e., a proportional odds model), a normal linear model for $g\left(\ln\left(M\right)|C,D\right)$, and a logit model for $h\left(Y|C,D,L,M\right)$, with $J = 2000$ simulations. The inverse probability weighting (IPW) estimates are based on a logit model for $f\left(D|C\right)$, an ordinal logit model for $q\left(L|C,D\right)$, and a normal linear model for $g\left(\ln\left(M\right)|C,D,L\right)$. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch4/figure_4-9`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/plowUse`.

We estimated the effects outlined previously using three different parametric approaches. First, we used regression-with-residuals (RWR), as described in Section 4.6.1. This approach involves fitting linear models for the mediator, outcome, and exposure-induced confounder. Aside from including an exposure-mediator interaction in the outcome regression, all models used to implement RWR were additive in the predictors.

Second, we applied the simulation approach described in Section 4.6.2. To implement this approach, we fit an ordinal logit model for the exposure-induced confounder, a log-normal model for the mediator, and a logit model for the outcome. As before, these models were all additive in the predictors, except for an exposure-mediator interaction included in the outcome model.

Finally, we estimated a third set of effects using inverse probability weights, following the procedures outlined in Section 4.6.3. To compute the weights, we fit a binary logit model for the probability of historical plow use, and we again opted to use an ordinal logit model for the degree of authoritarian versus democratic governance and a log-normal model for national income.

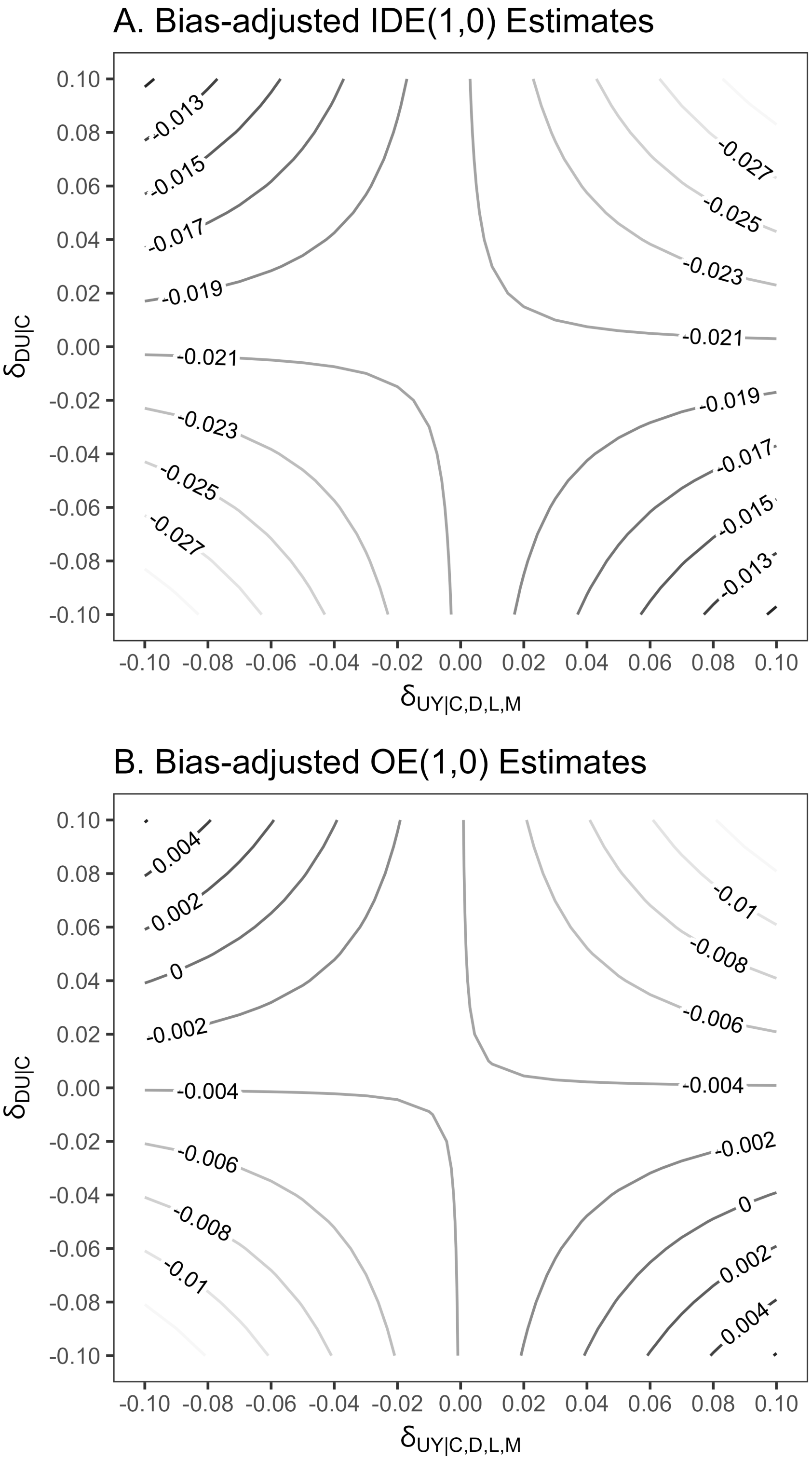
The results of this analysis are displayed in Figure 4.9, which presents point estimates and 95% confidence intervals based on the nonparametric bootstrap. Consistent with Alesina et al. (2013), we find evidence that historical plow use impacts female political participation indirectly through its effect on national income. This finding persists even after adjusting for exposure-induced confounding by the governance style of a country–be it authoritarian or democratic. For example, both RWR and the simulation approach yield estimates for the interventional indirect effect that suggest a 2 percentage point increase in female parliamentary representation attributable to the impact of historical plow cultivation on contemporary GDP.

However, we found weaker evidence to support a negative direct effect of plow use on female participation in politics. Specifically, while point estimates for the interventional and controlled direct effects are consistently negative, their confidence intervals are relatively wide and all span zero. This pattern of results was also consistent across estimation methods, indicating a degree of robustness to different modeling choices.

Figure 4.10 displays a set of contour graphs that describe the sensitivity of our estimates for the interventional direct and overall effects to exposure-outcome confounding by an unobserved variable. This concern arises due to our limited set of baseline covariates and the potential for numerous environmental, cultural, or contextual factors to confound the relationship between historical plow use and the evolution of contemporary gender roles.

To evaluate the sensitivity of our estimates, we posit the existence of a binary unobserved confounder U, and we additionally assume an invariant partial relationship of U with the outcome and a constant disparity in the prevalence of U across levels of the exposure. Under these assumptions, the bias in estimates for both the interventional direct effect and overall effect is given by $\text{Bias}\left(\widehat{IDE}\left(d,d^*\right)\right) = \text{Bias}\left(\widehat{OE}\left(d,d^*\right)\right) = \delta_{UY|C,D,L,M} \times \delta_{DU|C}$, where $\delta_{UY|C,D,L,M} = \mathbb{E}\left[Y|c,d,l,m,U=1\right] - \mathbb{E}\left[Y|c,d,l,m,U=0\right]$ and $\delta_{DU|C} = P\left(U=1|c,d\right) - P\left(U=1|c,d^*\right)$.

The contour graphs in Figure 4.10 plot bias-adjusted estimates across a range of values for the sensitivity parameters $\delta_{UY|C,D,M}$ and $\delta_{DU|C,M}$. These graphs indicate that our estimates for the interventional direct and overall effects are fairly stable under several different patterns of exposure-outcome confounding. Across all values for the sensitivity parameters, the bias-adjusted estimates remain substantively small and do not depart significantly from zero. Based on these findings, we conclude that if historical plow cultivation negatively affects the contemporary representation of women in government, its impact is likely quite modest and at least partially offset by the positive influence of plow use operating through higher national incomes. The data and code for replicating this analysis are accessible via hyperlinks provided in the footnotes accompanying Figures 4.9 and 4.10.



A. Bias-adjusted IDE(1,0) Estimates

B. Bias-adjusted OE(1,0) Estimates

Figure 4.10: Estimates of the Interventional Direct Effect and Overall Effect Adjusted for Bias due to Unobserved Exposure-Outcome Confounding.

Note: Estimates are based on RWR implemented with additive linear models for $\mathbb{E}[M|C,D]$ and $\mathbb{E}[L|C,D]$, and a linear model for $\mathbb{E}[Y|C,D,L,M]$ with a with $D \times M$ interaction. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch4/figure_4-10`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/plowUse`.

4.10 Summary

In this chapter, we examined methods for analyzing causal mediation in the presence of exposure-induced confounding. We began with a discussion of the natural effects decomposition and its limitations, where we showed that natural direct and indirect effects cannot be nonparametrically identified if there are any confounding variables influenced by the exposure. We then introduced a set of interventional direct and indirect effects that can still be identified when such confounders exist. These effects are similar to their natural counterparts but are formulated around randomized interventions on the mediator, making them identifiable under a less stringent set of conditions. Specifically, interventional effects can be nonparametrically identified provided that there are not any unobserved confounders distorting the relationships among the exposure, mediator, and outcome. Compared with the assumptions required for identifying natural effects, these conditions are considerably weaker, as they allow for the presence of exposure-induced confounders.

After defining our target estimands and explaining the conditions under which they can be identified, we described several approaches for estimating them, including regression-with-residuals (RWR), simulation via generalized linear models, and inverse probability weighting. As an extension of conventional linear regression analysis, RWR is best suited for applications with a continuous outcome and mediator, although it may also perform reasonably well when these variables are binary, ordinal, or counts. A key advantage of this approach over others is that it can easily accommodate multiple exposure-induced confounders.

In contrast, the simulation approach can be used with a much broader class of both linear and nonlinear models, allowing for more complex relationships and response distributions. Despite this flexibility, however, the simulation approach is more cumbersome to implement and susceptible to specification errors in applications with a large number of exposure-induced confounders.

Inverse probability weighting is also quite flexible in that it can be implemented with a wide variety of models, but this approach also becomes unwieldy in applications with more than one exposure-induced confounder. Moreover, its performance often suffers when the exposure, mediator, or confounder have a large number of values. Weighting is thus best suited for analyses in which all these variables are binary or otherwise discrete with only a few levels.

To maintain our focus and simplify the presentation of complex material, this chapter omitted discussions on certain advanced topics. These include methods for assessing how susceptible interventional effects are to biases from measurement error, as well as techniques for further partitioning these effects into components isolating different forms of interaction or moderation. Although these topics are relatively new areas of research and development, readers interested in them can refer to an emerging technical literature on sensitivity analysis for interventional effects (Lin et al. 2023; Park and Esterling 2021; Wodtke and Zhou 2020) and on effect decompositions involving interventional estimands (Nguyen et al. 2022; Wodtke and Zhou 2020; Wodtke et al. 2020, 2023).

Chapter 5

Mediation Analysis with Multiple Mediators

In the previous two chapters, we introduced methods for analyzing whether and to what extent a single mediator transmits the effect of an exposure on an outcome. We first considered settings where all potential confounders are measured prior to the exposure of interest or are otherwise not exposure-induced. We then examined scenarios where one or more exposure-induced variables may confound the mediator-outcome relationship. Specifically, in Chapter 3, we analyzed how unemployment may mediate the causal effect of college attendance on mental health, assuming that all potential confounders of the relationships between college attendance, unemployment, and mental health are measured before college attendance. In Chapter 4, we explored the mediating role of household income, treating unemployment as a potential exposure-induced confounder for the effect of income on mental health. In this context, we highlighted that the natural direct and indirect effects of income are not nonparametrically identified due to exposure-induced confounding. Instead, we analyzed a set of alternative estimands, known as interventional direct and indirect effects, which can still be nonparametrically identified and consistently estimated in the presence of exposure-induced confounders.

Yet, in the above example, both unemployment and household income may mediate the effect of college attendance on mental health, and we may want to quantify their respective contributions to the total effect of college attendance. In analyses with multiple mediators, analysts often attempt to examine one mediator at a time, without considering the potential for exposure-induced confounding. For example, we might consider applying the methods described in Chapter 3 twice: first with unemployment as the focal mediator and then with household income. This approach is problematic, however, because it ignores the potential for exposure-induced confounding when estimating the natural direct and indirect effects of household income, likely resulting in biased estimates. In fact, when the effects of interest involve multiple mediators that are *causally related*–that is, when one mediator affects another–the causal pathways operating through those mediators are not mutually exclusive, which precludes even a conceptual separation of their effects.

In this chapter, we first review two common approaches to analyzing multiple mediators: the one-mediator-at-a-time approach described above and an approach that examines all the mediators together (VanderWeele and Vansteelandt 2014; VanderWeele 2015). The first approach is only valid when the different mediators are *causally unrelated*–that is, when they do not influence each other–a condition that is strong, untestable, and unrealistic in many applications. The second approach allows causal relationships among the mediators; however, by treating all mediators collectively, it does not differentiate their individual contributions to the total effect.

Given the limitations of these two approaches, we introduce on a third approach that allows for causal relationships among the different mediators but still isolates their respective contributions to the total effect. With this approach, the average total effect is decomposed into a direct effect and a series of path-specific effects (PSEs; Avin et al. 2005), each of which reflects the net contribution of a specific mediator to the total effect above and beyond that of preceding mediators (Daniel et al. 2015; VanderWeele and Vansteelandt 2014; Zhou and Yamamoto 2022). We outline a set of assumptions that enable nonparametric identification of these PSEs and present their identification formulas. We then describe estimation strategies based on linear models, inverse probability weighting, and regression imputation, highlighting how these approaches connect with those outlined in previous chapters.

We continue to use data from the NLSY to illustrate key concepts and methods. In contrast to Chapters 3 and 4, we now consider both unemployment and household income as mediators that may transmit the effect of college attendance on depression, assuming that unemployment causally precedes income. Using the methods introduced in this chapter, we decompose the total effect of college attendance on depression into a direct effect, an indirect effect that operates through unemployment, and another indirect effect that operates solely through household income (and not via unemployment).

We also present a second empirical illustration using data from Brader et al. (2008), who conducted a survey experiment to evaluate the effect of negative media framing on support for immigration in the U.S. In their analyses of these data, the authors considered the mediating role of two different variables: beliefs about the economic costs of immigration and respondent anxiety. However, they analyzed these factors separately, under the assumption that anxiety is not affected by beliefs about economic costs. This assumption seems unlikely and appears inconsistent with the data (Imai and Yamamoto 2013). Thus, we treat these mediators as causally related and estimate their respective contributions to the total effect using PSEs.

Stata and R codes for implementing the analyses described in this chapter can be accessed at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch5`. Additionally, the footnotes accompanying each table and figure include hyperlinks to the specific scripts and data files used to generate their contents.

5.1 The One-Mediator-at-a-Time Approach

In the presence of multiple mediators, a common practice is to analyze each mediator separately, one at a time. For example, to unpack the causal effect of college attendance on depression, we might apply methods presented in Chapter 3 twice, first treating unemployment as the focal mediator and then treating household income as the focal mediator. The resulting estimates of natural direct and indirect effects would then be used to assess the mediating roles of unemployment and household income, respectively.

The one-mediator-at-a-time approach is valid if the identification assumptions for the natural direct and indirect effects–specifically, assumptions (c.i) to (c.vi) detailed in Section 3.3–hold for each mediator in question. The union of these assumptions over multiple mediators is highly restrictive because it requires that the mediators of interest are causally unrelated, meaning that they do not influence one another. In addition, these assumptions also do not allow unobserved or exposure-induced confounding for any of the mediator-mediator or mediator-outcome relationships. Figure 5.1 depicts a mediation model in which these conditions are satisfied. In this model, two mediators, denoted by M_1 and M_2, lie on separate causal paths

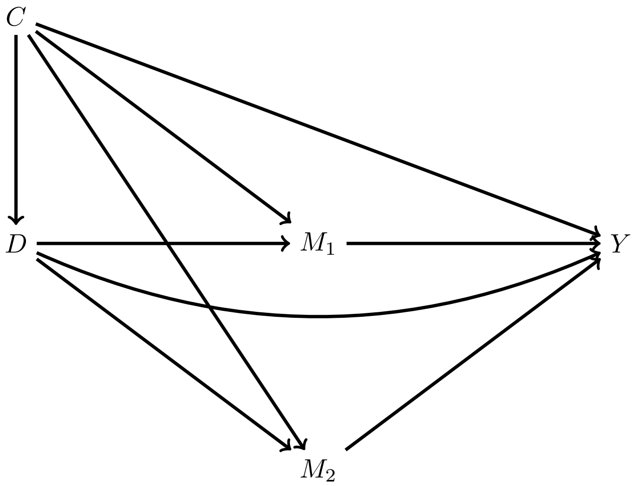


Figure 5.1: A Mediation Model with Baseline Confounding and Causally Unrelated Mediators.

Note: D denotes the exposure, M_1 and M_2 denote two mediators of interest, Y denotes the outcome, and C denotes a set of baseline confounders.

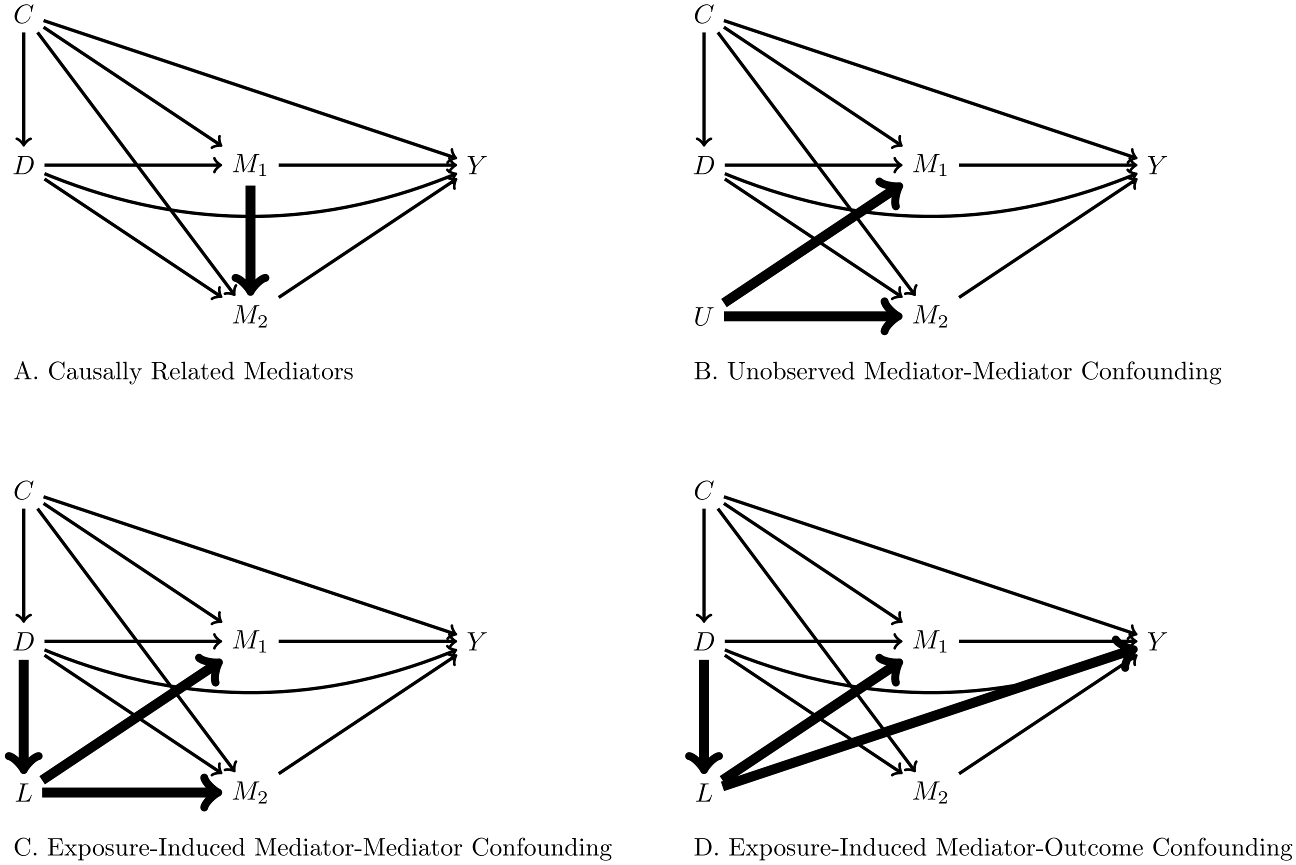
from the exposure to the outcome, with no causal path between them. Their relationships with the outcome are also unconfounded, given the exposure and baseline confounders.

Figure 5.2 presents several variations of Figure 5.1 where the one-mediator-at-a-time approach becomes problematic. Specifically, Panel A shows a mediation model where the two mediators, M_1 and M_2, are causally related, as indicated by the path $M_1 \to M_2$. In this scenario, because M_1 affects both M_2 and Y, and is itself affected by the exposure D, it acts as an exposure-induced confounder for the relationship between the second mediator and the outcome. Consequently, natural direct and indirect effects with respect to M_2 cannot be nonparametrically identified. Estimates of these effects from the one-mediator-at-a-time approach, which fails to account for exposure-induced confounding, would likely be biased.

In Panel B, an unobserved variable U affects both M_1 and M_2. Because of the causal paths $U \to M_1$ and $U \to M_2 \to Y$, U is an unobserved confounder for the relationship between M_1 and Y. Similarly, the paths $U \to M_2$ and $U \to M_1 \to Y$ also establish U as an unobserved confounder for the relationship between M_2 and Y. Thus, unobserved mediator-outcome confounding contaminates both the M_1-Y and M_2-Y relationships. As a result, natural effects cannot be nonparametrically identified for either M_1 or M_2, again rendering the one-mediator-at-a-time approach invalid.

The one-mediator-at-a-time approach would also be problematic in the presence of exposure-induced confounding of a mediator-mediator or mediator-outcome relationship, as illustrated in the last two panels of Figure 5.2. In Panel C, an exposure-induced variable L affects both M_1 and M_2. Echoing our earlier analysis of Panel B, the causal paths $L \to M_1 \to Y$ and $L \to M_2 \to Y$ establish that L is an exposure-induced confounder for both the M_1-Y and M_2-Y relationships. Consequently, the corresponding natural effects are not identified.

In Panel D, there is an exposure-induced confounder L only for the M_1-Y relationship. In this scenario, the one-mediator-at-a-time approach would remain valid for assessing the mediating role of M_2. However, natural direct and indirect effects with respect to M_1 are not identified due to exposure-induced confounding. To circumvent this problem, we might consider including L as another mediator for analysis using the one-mediator-at-a-time approach, but this too would be invalid, as the path from $L \to M_1$ implies that the



A. Causally Related Mediators

B. Unobserved Mediator-Mediator Confounding

C. Exposure-Induced Mediator-Mediator Confounding

D. Exposure-Induced Mediator-Outcome Confounding

Figure 5.2: Mediation Models in which the One-Mediator-at-a-Time Approach Becomes Problematic.

Note: D denotes the exposure, M_1 and M_2 denote two mediators, Y denotes the outcome, C denotes a set of observed baseline confounders, U denotes a set of unobserved confounders, and L denotes a set of exposure-induced confounders.

Table 5.1: Total, Direct, and Indirect Effects of College Attendance on CES-D Scores as Estimated from the One-Mediator-at-a-Time Approach Applied to the NLSY

Mediator	Estimand	Point Estimates and 95% CIs	
		Additive Linear Model (LinMod)	LinMod with $D \times M$ Interaction Term
M_1 (Unemployment)	$ATE\,(1,0)$	$-.079$ $[-.158, -.001]$	$-.083$ $[-.164, .000]$
	$NDE\,(1,0)$	$-.072$ $[-.150, .008]$	$-.078$ $[-.160, .006]$
	$NIE\,(1,0)$	$-.007$ $[-.017, .001]$	$-.005$ $[-.015, .001]$
M_2 (Household Income)	$ATE\,(1,0)$	$-.071$ $[-.151, .010]$	$-.078$ $[-.165, .011]$
	$NDE\,(1,0)$	$-.023$ $[-.107, .058]$	$-.038$ $[-.136, .062]$
	$NIE\,(1,0)$	$-.047$ $[-.064, -.032]$	$-.040$ $[-.068, -.018]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals, which were computed using the nonparametric bootstrap with $B = 2000$ replications. D denotes the exposure (college attendance) and M denotes the focal mediator. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch5/table_5-1`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

mediators of interest are causally related.

In sum, the one-mediator-at-a-time approach is only valid if the natural direct and indirect effects are nonparametrically identified for each mediator of interest. This condition requires that the different mediators are causally unrelated and that there is no unobserved or exposure-induced confounding for any of the mediator-mediator or mediator-outcome relationships. Under these assumptions, the natural effects for each mediator taken separately can be identified by the same formulas and estimated using the same strategies as discussed in Chapter 3, which include methods based on linear models, nonlinear models with simulation, and inverse probability weighting.

Applying the one-mediator-at-a-time approach to data from the NLSY, Table 5.1 reports estimates of the total, direct, and indirect effects of college attendance on depression, mediated by unemployment and household income. The estimates in the first column are based on additive linear models, equivalent to Equations 3.27 and 3.28. The estimates in the second column come from an additive linear model for the mediator and then a linear model for the outcome that includes an exposure-mediator interaction term, as in Equations 3.30 and 3.31. When fitting these models, we include the same set of baseline confounders as described in Chapter 3. Thus, the point estimates shown in the upper panel simply reproduce those reported in Table 3.2, which indicate that unemployment plays a minimal role in transmitting the effect of college attendance on depression. By contrast, our estimates of natural direct and indirect effects in the lower panel suggest a substantial mediating role for household income. For example, estimates from the more flexible specification that includes an exposure-mediator interaction suggest that about half of the total effect ($0.04/0.078 = 51\%$) operates through household income. Links to the code and data used for this analysis are provided in the table footnote.

Nevertheless, as we cautioned earlier, the one-mediator-at-a-time approach relies on a set of strong assumptions, which, if violated, can lead to biased estimates of natural effects for one or more of the mediators. In our analysis of the NLSY, the assumption that the different mediators are causally unrelated is particularly dubious, as unemployment (M_1) almost certainly has an adverse effect on household income (M_2). The assumption that M_1 does not affect M_2 is not directly testable without assuming away unobserved confounding of the M_1-M_2 relationship. However, if all the assumptions that justify the one-mediator-at-a-time approach are satisfied, M_1 should be statistically independent of M_2, conditional on the exposure D and baseline confounders C. Conversely, M_1 and M_2 would not be independent, given D and C, under any of the four scenarios depicted in Figure 5.2. Thus, to assess the joint validity of these assumptions, we can test whether M_1 and M_2 are conditionally independent.

A straightforward way to implement this test is to regress M_2 on C, D, and M_1 (Imai and Yamamoto 2013). If all required assumptions hold and the regression for M_2 is correctly specified, the coefficient on M_1 should be zero. Alternatively, rejecting the null hypothesis that this coefficient is zero would indicate that the two mediators are either causally related or their relationship is confounded by unobserved variables. In either case, the one-mediator-at-a-time approach is inappropriate.

When we apply this test to the NLSY, we find that the coefficient on unemployment is negative (-.44), and the associated p-value provides strong evidence against the possibility that it is equal to zero in the target population ($p < 0.001$). This suggests that unemployment and household income are either causally related or jointly influenced by unobserved factors. Since both scenarios violate the assumptions required of the one-mediator-at-a-time approach, the estimates for the direct and indirect effects in Table 5.1 are likely biased. In the next section, we outline an alternative approach that can accommodate both of these challenges–that is, causal dependence among mediators and unobserved mediator-mediator confounding–by analyzing the explanatory role of the different mediators taken as a whole.

5.2 The Multiple-Mediators-as-a-Whole Approach

In this section, we describe an alternative approach to analyzing multiple mediators that treats all of them as a whole (VanderWeele 2015; VanderWeele and Vansteelandt 2014). This approach is less ambitious than the one-mediator-at-a-time approach in that it does not attempt to isolate the mediating roles of each variable taken separately. Nevertheless, it remains useful, especially when different mediators can be conceptualized as alternative indicators of a broader theoretical construct. For example, to evaluate the role of unemployment (M_1) and household income (M_2) in transmitting the effect of college attendance on depression, this approach combines the two variables to form a multivariate mediator denoted by $\mathbf{M} = (M_1, M_2)$, which can be interpreted as a vector-valued measure of "socioeconomic conditions." It then focuses solely on direct and indirect effects with respect to $\mathbf{M}$. These effects can be formally defined as follows:

$$\begin{aligned} NDE_{\mathbf{M}}(d, d^*) &= \mathbb{E}\left[Y\left(d, \mathbf{M}(d^*)\right) - Y\left(d^*, \mathbf{M}(d^*)\right)\right] \\ NIE_{\mathbf{M}}(d, d^*) &= \mathbb{E}\left[Y\left(d, \mathbf{M}(d)\right) - Y\left(d, \mathbf{M}(d^*)\right)\right], \end{aligned} \tag{5.1}$$

which we refer to as *multivariate natural direct and indirect effects*. The $NDE_{\mathbf{M}}(d, d^*)$ captures an effect of the exposure on the outcome that does not operate through any of the mediators included in $\mathbf{M}$, while the $NIE_{\mathbf{M}}(d, d^*)$ captures an effect transmitted specifically through this set of mediators considered altogether.

The multiple-mediators-as-a-whole approach requires the identification assumptions for natural direct and indirect effects–specifically, assumptions (c.i) to (c.vi) from Section 3.3–to hold for all mediators taken as a single entity. These assumptions are encoded in Figure 5.3, which depicts a model with two mediators, M_1 and M_2, combined to form a vector-valued mediator $\mathbf{M} = (M_1, M_2)$. This model is much less restrictive

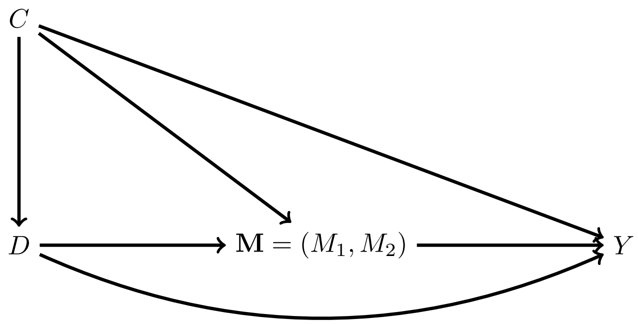


Figure 5.3: A Mediation Model that Analyzes Multiple Mediators Collectively.

Note: D denotes the exposure, M_1 and M_2 denote two mediators of interest, Y denotes the outcome, and C denotes a set of baseline confounders.

than the mediation model in Figure 5.1, as it can accommodate all the scenarios depicted in Figure 5.2 that would invalidate the one-mediator-at-a-time approach.

First, because it focuses on direct and indirect effects with respect to $\mathbf{M}$, the model remains agnostic about any causal relationships among the components of this vector, thus allowing the different mediators in $\mathbf{M}$ to influence each other, as in Panel A of Figure 5.2. Second, the model permits the presence of unobserved variables that affect both M_1 and M_2, as in Panel B of Figure 5.2, because such variables would not confound any of the relationships between the exposure and outcome, exposure and mediators, or mediators and outcome. For the same reason, it also allows for exposure-induced variables that affect both M_1 and M_2, as in Panel C of Figure 5.2. Finally, in situations where an exposure-induced variable L is thought to confound any of the mediator-outcome relationships, as depicted in Panel D of Figure 5.2, L can simply be incorporated into the vector of mediators $\mathbf{M}$. In such cases, assumptions (c.i) to (c.vi) will remain valid with respect to $\mathbf{M} = (L, M_1, M_2)$,and the multivariate direct and indirect effects will capture the mediating role of all these variables operating together.

Under assumptions (c.i) to (c.vi) with respect to $\mathbf{M}$, the $NDE_{\mathbf{M}}(d, d^*)$ and $NIE_{\mathbf{M}}(d, d^*)$ can be identified and consistently estimated with methods similar to those discussed in Chapter 3. These include estimators based on linear models, nonlinear models with simulation, and inverse probability weighting.

Specifically, the weighting estimator described in Section 3.5.3 can be directly applied to estimate multivariate direct and indirect effects. For this application, we simply need to include all the elements of $\mathbf{M}$ as predictors when fitting the model for the exposure conditional on the baseline confounders and mediator(s). Otherwise, the method is implemented exactly as outlined previously in Chapter 3.

The other estimators from Chapter 3 require more elaborate modifications when targeting the $NDE_{\mathbf{M}}(d, d^*)$ and $NIE_{\mathbf{M}}(d, d^*)$. To extend the approach based on linear models, initially described in Section 3.5.1, for multivariate direct and indirect effects, we need to fit separate models for each mediator and include all the mediators as predictors in the model for the outcome. Let $\mathbf{M} = (M_1, M_2, \ldots, M_K)$ denote a multivariate mediator with K components, and consider first the following set of additive linear models:

$$\mathbb{E}\left[M_k|c,d\right] = \beta_{0k} + \beta_{1k}^T c + \beta_{2k} d \quad \text{for } k = 1, 2, \ldots, K \tag{5.2}$$

$$\mathbb{E}\left[Y|c,d,\mathbf{m}\right] = \gamma_0 + \gamma_1^T c + \gamma_2 d + \sum_{k=1}^{K} \gamma_{3k} m_k, \tag{5.3}$$

where $\mathbb{E}\left[M_k|c,d\right]$ denotes the conditional mean of M_k given the exposure and baseline confounders and $\mathbb{E}\left[Y|c,d,\mathbf{m}\right]$ denotes the conditional mean of the outcome given the exposure, baseline confounders, and all the mediators in $\mathbf{M}$. If assumptions (c.i) to (c.vi) are satisfied with respect to $\mathbf{M}$ and Equations 5.2 and 5.3 are correctly specified, then the direct and indirect effects of interest are given by $NDE_{\mathbf{M}}\left(d,d^*\right) = \gamma_2\left(d - d^*\right)$ and $NIE_{\mathbf{M}}\left(d,d^*\right) = \left(\sum_{k=1}^{K} \beta_{2k}\gamma_{3k}\right)\left(d - d^*\right)$, respectively, and the average total effect is given by their sum. Moreover, because there is no interaction between the exposure and mediator in the outcome model, a multivariate analogue of the controlled direct effect is equal to the $NDE_{\mathbf{M}}\left(d,d^*\right)$–that is, $CDE\left(d,d^*,\mathbf{m}\right) = \gamma_2\left(d - d^*\right)$ as well.

Consistent estimators for these effects can be constructed by fitting Equations 5.2 and 5.3 using the method of OLS and then by substituting their coefficient estimates into the appropriate parametric expressions. Specifically, a set of consistent estimators can be constructed as follows:

$$\begin{aligned}
\widehat{NDE}\left(d,d^*\right)^{lm} &= \widehat{CDE}\left(d,d^*,\mathbf{m}\right)^{lm} = \hat{\gamma}_2\left(d - d^*\right) \\
\widehat{NIE}\left(d,d^*\right)^{lm} &= \left(\sum_{k=1}^{K} \hat{\beta}_{2k}\hat{\gamma}_{3k}\right)\left(d - d^*\right) \\
\widehat{ATE}\left(d,d^*\right)^{lm} &= \left(\hat{\gamma}_2 + \sum_{k=1}^{K} \hat{\beta}_{2k}\hat{\gamma}_{3k}\right)\left(d - d^*\right),
\end{aligned} \tag{5.4}$$

where the “hats” distinguish estimators from estimands, as before, and where the “lm” superscript indicates that these estimators are based on additive linear models fit by least squares.

The expressions in Equation 5.4 are based on highly restrictive models, and thus they are prone to misspecification bias. In particular, the models on which these estimators are based do not allow the effects of the exposure and mediator to interact or to vary across levels of the baseline confounders. If the exposure and mediator interact or their effects are moderated by the confounders, then the estimators based on Equations 5.2 and 5.3 will be biased and inconsistent, as the relationship of these variables with each other and with the outcome is constrained to be additive, when in fact it is not.

Fortunately, as in the case of a univariate mediator, estimation with linear models can be easily adapted to accommodate interaction effects between the exposure and the different mediators in $\mathbf{M}$. Consider next the following set of linear models for the mediator and outcome:

$$\mathbb{E}\left[M_k|c,d\right] = \beta_{0k} + \beta_{1k}^T c^{\perp} + \beta_{2k} d \quad \text{for } k = 1, 2, \ldots, K \tag{5.5}$$

$$\mathbb{E}\left[Y|c,d,\mathbf{m}\right] = \gamma_0 + \gamma_1^T c^{\perp} + \gamma_2 d + \sum_{k=1}^{K} m_k\left(\gamma_{3k} + \gamma_{4k} d\right). \tag{5.6}$$

These models are nearly identical to Equations 5.2 and 5.3 above except for two important differences. First, in the outcome model, we have now included all exposure-mediator interaction effects, given by the coefficients γ_{4k}. Second, in both models, we have centered the baseline confounders around their sample means, where $c^{\perp} = c - \bar{C}$.

If assumptions (c.i) to (c.vi) are satisfied with respect to $\mathbf{M}$ and Equations 5.5 and 5.6 are correctly specified, then the multivariate natural effects are given by $NDE_{\mathbf{M}}(d, d^*) = \left(\gamma_2 + \sum_{k=1}^{K} \gamma_{4k} (\beta_{0k} + \beta_{2k} d^*)\right)(d - d^*)$ and $NIE_{\mathbf{M}}(d, d^*) = \left(\sum_{k=1}^{K} \beta_{2k} (\gamma_{3k} + \gamma_{4k} d)\right)(d - d^*)$, while the total effect is given by their sum. Moreover, because the outcome model incorporates interaction terms between the exposure and mediators, the parametric expression for the controlled direct effect now differs from that for the natural direct effect. In particular, the controlled direct effect is given by $CDE(d, d^*, \mathbf{m}) = \left(\gamma_2 + \sum_{k=1}^{K} \gamma_{4k} m_k\right)(d - d^*)$, where $\mathbf{m} = (m_1, m_2, \ldots, m_K)$ denotes the levels of each mediator M_k at which this effect is evaluated.

The models in Equations 5.5 and 5.6 are more flexible in that they allow for exposure-mediator interaction effects on the outcome, but they still constrain the effects of the exposure and mediator to be invariant across levels of the baseline confounders. In many social science applications, where effect heterogeneity is endemic, this constraint may also be unrealistic.

Estimation with linear models can be adapted even further to accommodate effect moderation by the baseline confounders. Consider now the following set of models, which allow the effects of the exposure and mediator to vary across levels of the confounders:

$$\mathbb{E}[M_k|c, d] = \beta_{0k} + \beta_{1k}^T c^{\perp} + d\left(\beta_{2k} + \beta_{3k}^T c^{\perp}\right) \quad \text{for } k = 1, 2, \ldots, K \tag{5.7}$$

$$\mathbb{E}[Y|c, d, \mathbf{m}] = \gamma_0 + \gamma_1^T c^{\perp} + \gamma_2 d + \sum_{k=1}^{K} m_k (\gamma_{3k} + \gamma_{4k} d) + c^{\perp} \sum_{k=1}^{K} \left(\gamma_{5k}^T d + m_k \left(\gamma_{6k}^T + \gamma_{7k}^T d\right)\right). \tag{5.8}$$

These models are similar to those in Equations 5.5 and 5.6 above except for a few important differences. Specifically, in each mediator model, we have additionally incorporated interactions between the exposure and baseline confounders, where β_{3k}^T is a transposed vector of coefficients multiplying all two-way interactions of the exposure with the mean-centered confounders. Additionally, in the model for the outcome, we have incorporated interactions of the exposure and mediator with the baseline confounders, where the coefficient vectors denoted by $\{\gamma_{5k}^T, \gamma_{6k}^T, \gamma_{7k}^T\}$ allow the joint effects of both the exposure and each mediator to differ across levels of the baseline confounders.

Provided that all of these new interaction terms are constructed after mean-centering the baseline confounders, and not with their original values, the parametric expressions for the total, direct, and indirect effects of interest are exactly the same as those based on Equations 5.5 and 5.6 above. That is, the parametric expressions for the effects of interest are unchanged by incorporating confounder-by-exposure and confounder-by-mediator interactions, as long as these interactions are constructed with the mean-centered confounders (Wodtke et al. 2020; Wodtke and Zhou 2020). Adding these terms merely relaxes modeling restrictions on the joint distribution of the observed data by allowing for certain patterns of effect moderation across levels of the baseline confounders, potentially mitigating bias due to misspecification.

Under models resembling those in Equations 5.5 to 5.6, consistent estimators for the effects of interest can be constructed by fitting these models to sample data using OLS and then by substituting their coefficient estimates into the parametric expressions provided above. Specifically, a set of consistent estimators can be constructed from the fitted models as follows:

$$\begin{aligned}
\widehat{NDE}\left(d, d^{*}\right)^{lmi} &= \left(\hat{\gamma}_{2} + \sum_{k=1}^{K} \hat{\gamma}_{4k}\left(\hat{\beta}_{0k} + \hat{\beta}_{2k} d^{*}\right)\right)\left(d - d^{*}\right) \\
\widehat{NIE}\left(d, d^{*}\right)^{lmi} &= \left(\sum_{k=1}^{K} \hat{\beta}_{2k}\left(\hat{\gamma}_{3k} + \hat{\gamma}_{4k} d\right)\right)\left(d - d^{*}\right) \\
\widehat{ATE}\left(d, d^{*}\right)^{lmi} &= \left(\hat{\gamma}_{2} + \sum_{k=1}^{K} \hat{\gamma}_{4k}\left(\hat{\beta}_{0k} + \hat{\beta}_{2k} d^{*}\right) + \sum_{k=1}^{K} \hat{\beta}_{2k}\left(\hat{\gamma}_{3k} + \hat{\gamma}_{4k} d\right)\right)\left(d - d^{*}\right) \\
\widehat{CDE}\left(d, d^{*}, \mathbf{m}\right)^{lmi} &= \left(\hat{\gamma}_{2} + \sum_{k=1}^{K} \hat{\gamma}_{4k} m_{k}\right)\left(d - d^{*}\right), \qquad (5.9)
\end{aligned}$$

where the “lmi” superscript indicates that these estimators are based on linear models with exposure-mediator interactions and possibly interactions of the exposure and mediator with the mean-centered confounders as well.

The simulation estimator introduced in Section 3.5.2 can also be extended to accommodate multiple mediators. However, because it involves modeling the conditional distribution of the mediator given the exposure and baseline confounders, it is somewhat more challenging to implement when the mediator is multivariate, in which case the analyst would need to specify and fit a model for the joint distribution of $\mathbf{M} = (M_1, M_2, \ldots, M_K)$ given the exposure D and baseline confounders C. For example, if $\mathbf{M}$ contained only continuous variables, the analyst might consider a multivariate normal model for $\mathbf{M}$, where the conditional means, conditional variances, and covariances between the elements of $\mathbf{M}$ are all specified as parametric functions of the exposure and baseline confounders.

Alternatively, the analyst could also model the joint distribution of $\mathbf{M}$ by fitting a series of GLMs for the conditional distribution of each mediator sequentially, using the confounders, exposure, and all preceding mediators as predictors (Zhou and Wodtke 2024). For example, if $\mathbf{M} = (M_1, M_2)$, we could first fit a model for the conditional distribution of M_1 given C and D, and then we could fit a second model for the conditional distribution of M_2 given C, D, and M_1. These models can then be used to simulate potential values of the multivariate mediator under different levels of the exposure, following procedures similar to those outlined in Section 4.6.2 for implementing the simulation estimator with multiple exposure-induced confounders.

Table 5.2 presents estimates for multivariate natural effects from the NLSY, computed using linear models as well as inverse probability weighting. Specifically, it reports estimates for the total, direct, and indirect effects of college attendance on depression, as mediated by unemployment and income jointly. The estimates in the first column are based on additive linear models equivalent to Equations 5.2 and 5.3. The estimates in the second column come from an additive linear model for each mediator and then a linear model for the outcome that includes all exposure-mediator interaction terms, as in Equations 5.5 and 5.6. The third column shows estimates based on inverse probability weighting. Links to the code and data used for this analysis are provided in the table footnote.

Consistent with the results reported in previous chapters, inverse probability weighting suggests a somewhat larger total effect of college attendance on depression compared with the estimates based on linear models. Nevertheless, across all three approaches to estimation, the results suggest a substantial role for “socioeconomic conditions,” as measured by unemployment and household income together, in mediating the effect of college attendance on mental health.

In sum, the multiple-mediators-as-a-whole approach is a useful strategy for assessing mediation while

Table 5.2: Total, Direct, and Indirect Effects of College Attendance on CES-D Scores as Estimated from the Multiple-Mediators-as-a-Whole Approach with the NLSY

Estimand	Point Estimates and 95% CIs		
	Additive Linear Model (LinMod)	LinMod with $D \times M$ Interaction Terms	Inverse Probability Weighting
$ATE\,(1,0)$	$-.070$ $[-.152, .011]$	$-.077$ $[-.163, .009]$	$-.167$ $[-.265, -.054]$
$NDE_{\mathbf{M}}\,(d, d^{*})$	$-.023$ $[-.104, .059]$	$-.037$ $[-.131, .060]$	$-.100$ $[-.232, .052]$
$NIE_{\mathbf{M}}\,(d, d^{*})$	$-.047$ $[-.066, -.029]$	$-.040$ $[-.069, -.016]$	$-.068$ $[-.147, -.009]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals, which were computed using the nonparametric bootstrap with $B = 2000$ replications. D denotes the exposure (college attendance) and $\mathbf{M}$ denotes the (vector-valued) mediator. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch5/table_5-2`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

circumventing the stringent assumptions associated with the one-mediator-at-a-time approach. It can be easily implemented with estimators based on inverse probability weighting and linear models for the outcome and the mediators. It can also be implemented using simulation methods and GLMs, although this estimation strategy is somewhat more cumbersome as it requires modeling the joint distribution of the mediators.

Compared to the one-mediator-at-a-time approach, the multiple-mediators-as-a-whole approach adopts a more conservative aim by focusing solely on multivariate direct and indirect effects, analyzing the explanatory role of all mediators collectively. Despite its comparatively modest ambitions, this approach can still provide valuable insights into the potential mechanisms driving a total effect. For example, our empirical analysis of the NLSY suggests that unemployment and household income jointly mediate about half of the total effect of college attendance on depression. However, distinguishing the individual contributions of each mediator to the total effect is impossible without additional assumptions.

5.3 Effect Decomposition with Multiple Mediators

In the previous sections, we explored two approaches for analyzing multiple mediators: the one-mediator-at-a-time approach and the multiple-mediators-as-a-whole approach. The former is suitable only when the mediators are causally unrelated and their relationships with each other and the outcome are unconfounded. In contrast, the latter approach accommodates causally related mediators and various forms of mediator-mediator and mediator-outcome confounding. However, by treating all mediators as a whole, this approach is unable to isolate each mediator's individual contribution to the total effect of interest.

Given these limitations, we now introduce a third approach that can accommodate causally related mediators and isolate their respective contributions to the total effect. Specifically, in this section, we show that with multiple mediators–whether they are causally related or not–the average total effect can be decomposed into a direct effect and a series of path-specific effects (PSEs), each of which quantifies the net contribution of a specific mediator to the total effect (Daniel et al. 2015; VanderWeele et al. 2014). We next introduce a set of assumptions that enable nonparametric identification of this decomposition and provide the corresponding identification formulas. Drawing on these formulas, we then discuss several different estimation strategies. Although this approach can accommodate an arbitrary number of mediators, we focus mainly on applications with two mediators for illustrative purposes. In Section 5.7 below, we generalize these methods to accommodate $K (\geq 2)$ causally ordered mediators.

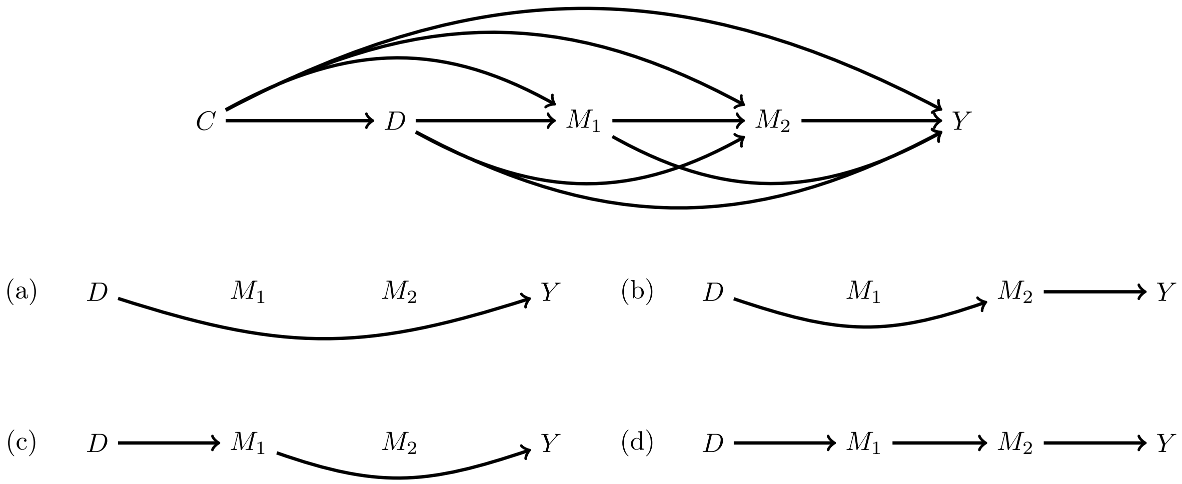


Figure 5.4: A Mediation Model with Two Causally Ordered Mediators.
Note: D denotes the exposure, Y denotes the outcome of interest, C denotes a vector of baseline confounders, and M_1 and M_2 denote two causally ordered mediators. The confounding arcs between C and each of the other nodes are omitted in subgraphs (a)-(d).

As before, let D denote an exposure, Y an outcome of interest, and C a vector of baseline confounders. Without loss of generality, first consider the scenario in which two mediators, M_1 and M_2, lie on causal paths from the exposure to the outcome. In addition, assume that M_1 causally precedes M_2, such that M_1 can influence M_2 but not vice versa.[1] In the NLSY, D represents college attendance, Y represents depression at midlife, M_1 represents unemployment, and M_2 represents household income.

A mediation model illustrating the hypothesized relationships among these variables is depicted in the top panel of Figure 5.4. In this model, four causal paths connect the exposure to the outcome, as shown in the lower panels of the figure: (a) $D \to Y$; (b) $D \to M_2 \to Y$; (c) $D \to M_1 \to Y$; and (d) $D \to M_1 \to M_2 \to Y$. If the mediators were causally unrelated, the last path would not exist. In this case, the total effect of D on Y could be neatly partitioned into an indirect effect operating through M_1, as represented by the $D \to M_1 \to Y$ path; an indirect effect operating through M_2, as represented by the $D \to M_2 \to Y$ path; and a direct effect that does not operate through either M_1 or M_2, as represented by the $D \to Y$ path (Imai and Yamamoto 2013). However, in the more common scenario where M_1 and M_2 are causally related, it is not possible to partition the mediating effects of M_1 and M_2 into entirely separate components, because part of the total effect operates through both M_1 and M_2 together, as represented by the $D \to M_1 \to M_2 \to Y$ path.

5.3.1 Path-specific Effects

To resolve this dilemma, PSEs partition the total effect into components that capture the contribution of each specific mediator, net of the contributions of other mediators that precede it in causal order. To formally

[1] Note that M_1 and M_2 can each be multivariate and that the causal relationships among the component variables can be left unspecified, as long as all the elements of M_1 causally precede the elements of M_2.

define these effects, we again rely on potential outcomes notation. Let $Y(d)$ denote the conventional potential outcome when the exposure is set to level d, and let $Y(d, m_1, m_2)$ denote the joint potential outcome when the exposure is set to d, the first mediator is set to level m_1, and the second mediator is set to level m_2. Similarly, $M_2(d, m_1)$ denotes the potential value of the second mediator if an individual experienced exposure d and level m_1 of the first mediator, while $M_1(d)$ denotes the potential value of the first mediator under exposure d. Finally, let $Y(d, M_1(d^*), M_2(d^*, M_1(d^*)))$ represent a nested, cross-world potential outcome, where the exposure is set to level d but the mediators are set to their values that would have arisen naturally had an individual experienced level d^* of the exposure instead. Because $Y(d)$ denotes the potential outcome when the exposure is set to level d and the mediators take on their natural values under this same exposure, $Y(d) = Y(d, M_1(d), M_2(d, M_1(d)))$ by definition.

With this notation, the average total effect of D on Y can be decomposed as follows:

$$\begin{aligned}
\mathbb{E}[Y(d) - Y(d^*)] &= \mathbb{E}[Y(d, M_1(d), M_2(d, M_1(d))) - Y(d^*, M_1(d^*), M_2(d^*, M_1(d^*)))] \\
&= \underbrace{\mathbb{E}[Y(d, M_1(d^*), M_2(d^*, M_1(d^*))) - Y(d^*, M_1(d^*), M_2(d^*, M_1(d^*)))]}_{D \to Y} \\
&\quad + \underbrace{\mathbb{E}[Y(d, M_1(d^*), M_2(d, M_1(d^*))) - Y(d, M_1(d^*), M_2(d^*, M_1(d^*)))]}_{D \to M_2 \to Y} \\
&\quad + \underbrace{\mathbb{E}[Y(d, M_1(d), M_2(d, M_1(d))) - Y(d, M_1(d^*), M_2(d, M_1(d^*)))]}_{D \to M_1 \to Y;\, D \to M_1 \to M_2 \to Y} \\
&= PSE_{D \to Y}(d, d^*) + PSE_{D \to M_2 \to Y}(d, d^*) + PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*), \qquad (5.10)
\end{aligned}$$

The three terms in this decomposition represent the PSEs corresponding to the $D \to Y$, $D \to M_2 \to Y$, and $D \to M_1 \rightsquigarrow Y$ paths, where the squiggly arrow represents the combination of multiple paths emanating from M_1. Specifically, the first term, denoted by $PSE_{D \to Y}(d, d^*)$, quantifies the effect of the exposure on the outcome if both mediators were set to their values that would have arisen naturally under exposure d^*. It captures a direct effect of the exposure that does not operate through either mediator, as represented by the path $D \to Y$. The second term, $PSE_{D \to M_2 \to Y}(d, d^*)$, captures an indirect effect of the exposure on the outcome that operates through the second mediator but not the first, as represented by the path $D \to M_2 \to Y$. The last term, $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$, captures an indirect effect of the exposure on the outcome operating through the first mediator. It represents the combination of the paths $D \to M_1 \to Y$ and $D \to M_1 \to M_2 \to Y$, which together are denoted by $D \to M_1 \rightsquigarrow Y$.

Although four causal paths connect D to Y, Equation 5.10 partitions the total effect into only three components: $PSE_{D \to Y}(d, d^*)$, $PSE_{D \to M_2 \to Y}(d, d^*)$, and $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$. The last component, $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$, captures the effects transmitted along both the $D \to M_1 \to Y$ and the $D \to M_1 \to M_2 \to Y$ paths. It reflects the overall mediating effect of M_1, part of which may operate through its influence on M_2. By contrast, the second component of the decomposition, $PSE_{D \to M_2 \to Y}(d, d^*)$, captures only the path $D \to M_2 \to Y$, but not the path $D \to M_1 \to M_2 \to Y$. Thus, it should not be interpreted as an overall mediating effect of M_2. Rather, it reflects an "independent" mediating effect of M_2 above and beyond that due to M_1.

In the NLSY, the $PSE_{D \to Y}(d, d^*)$ reflects the direct effect of college attendance on depression–that is, the component of the total effect that does not operate through unemployment or income. The $PSE_{D \to M_2 \to Y}(d, d^*)$ reflects the effect of college attendance operating only through household income. And the $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$ reflects the effect of college attendance operating through unemployment, including

Table 5.3: PSEs that Compose Natural Direct and Indirect Effects in the Presence of Two Causally Related Mediators.

	NDE for M_2	NIE for M_2
NDE for M_1	PSE for $D \to Y$	PSE for $D \to M_2 \to Y$
NIE for M_1	PSE for $D \to M_1 \to Y$	PSE for $D \to M_1 \to M_2 \to Y$

any influence of unemployment on depression that is transmitted through its impact on household income.

5.3.2 Connections with Natural Direct and Indirect Effects

Consider now the natural direct and indirect effects via each of the mediators taken separately. For example, natural direct and indirect effects with respect to M_2 can be expressed through the following decomposition of the total effect (Imai and Yamamoto 2013):

$$\begin{aligned}\mathbb{E}\left[Y\left(d\right)-Y\left(d^{*}\right)\right] &= \underbrace{\mathbb{E}\left[Y\left(d,M_1\left(d\right),M_2\left(d^{*},M_1\left(d^{*}\right)\right)\right)-Y\left(d^{*},M_1\left(d^{*}\right),M_2\left(d^{*},M_1\left(d^{*}\right)\right)\right)\right]}_{D\to Y;D\to M_1\to Y} \\ &\quad+\underbrace{\mathbb{E}\left[Y\left(d,M_1\left(d\right),M_2\left(d,M_1\left(d\right)\right)\right)-Y\left(d,M_1\left(d\right),M_2\left(d^{*},M_1\left(d^{*}\right)\right)\right)\right]}_{D\to M_2\to Y;D\to M_1\to M_2\to Y} \\ &= NDE_{M_2}(d,d^{*})+NIE_{M_2}(d,d^{*}),\end{aligned} \tag{5.11}$$

In this decomposition, the $NIE_{M_2}(d,d^{*})$ represents the effect of the exposure on the outcome operating through the second mediator, including any influence of M_1 on M_2. Similarly, the $NDE_{M_2}(d,d^{*})$ reflects the effect of the exposure that does not operate through M_2, which may include the mediating influence of M_1 that operates independently of M_2. A critical limitation of this decomposition, however, is that neither the $NDE_{M_2}(d,d^{*})$ nor the $NIE_{M_2}(d,d^{*})$ can be nonparametrically identified because M_1 is an exposure-induced confounder of the relationship between M_2 and Y, as outlined previously.

By contrast, the PSEs defined in Equation 5.10 are aligned with the following alternative decomposition into natural direct and indirect effects through M_1:

$$\begin{aligned}\mathbb{E}\left[Y\left(d\right)-Y\left(d^{*}\right)\right] &= \underbrace{\mathbb{E}\left[Y\left(d,M_1\left(d^{*}\right),M_2\left(d,M_1\left(d^{*}\right)\right)\right)-Y\left(d^{*},M_1\left(d^{*}\right),M_2\left(d^{*},M_1\left(d^{*}\right)\right)\right)\right]}_{D\to Y;D\to M_2\to Y} \\ &\quad+\underbrace{\mathbb{E}\left[Y\left(d,M_1\left(d\right),M_2\left(d,M_1\left(d\right)\right)\right)-Y\left(d,M_1\left(d^{*}\right),M_2\left(d,M_1\left(d^{*}\right)\right)\right)\right]}_{D\to M_1\to Y;D\to M_1\to M_2\to Y} \\ &= NDE_{M_1}\left(d,d^{*}\right)+NIE_{M_1}\left(d,d^{*}\right),\end{aligned} \tag{5.12}$$

where the two terms represent the natural effects with respect to the first mediator, rather than the second. A comparison of Equation 5.12 with Equation 5.10 reveals that $NIE_{M_1}\left(d,d^{*}\right)=PSE_{D\to M_1\rightsquigarrow Y}\left(d,d^{*}\right)$ and $NDE_{M_1}\left(d,d^{*}\right) = PSE_{D\to Y}\left(d,d^{*}\right)+PSE_{D\to M_2\to Y}\left(d,d^{*}\right)$. Thus, the decomposition of the total effect into PSEs, as given in Equation 5.10, isolates the effect of the exposure that operates through M_1 ($NIE_{M_1}\left(d,d^{*}\right)=PSE_{D\to M_1\rightsquigarrow Y}\left(d,d^{*}\right)$) and then further decomposes the natural direct effect with respect to M_1 into an effect operating through M_2 but not through M_1 ($PSE_{D\to M_2\to Y}\left(d,d^{*}\right)$) and an effect operating neither through M_1 nor M_2 ($PSE_{D\to Y}\left(d,d^{*}\right)$).

Table 5.3 summarizes how the PSEs relate to natural direct and indirect effects with respect to M_1 and M_2, illustrating how PSEs represent a further decomposition of natural effects. The table also clarifies that if the mediators are causally unrelated–specifically, if the causal path $A \to M_1 \to M_2 \to Y$ does not exist–then the natural indirect effect with respect to M_1 is equivalent to a PSE involving only the first mediator, while the natural indirect effect with respect to M_2 amounts to a PSE involving only the second mediator. This illustrates that the common practice of analyzing one mediator at a time is just a special case of the decomposition into PSEs. Thus, even in scenarios where analysts are willing to assume that the different mediators are causally unrelated, the effect decomposition given in Equation 5.10 can still be applied. And in this situation, the PSEs through M_1 and M_2 can be equivalently interpreted as the natural indirect effects of these mediators.

5.4 Nonparametric Identification

Recall that a DAG represents a nonparametric structural equation model (Pearl 2009). The graph in the upper panel of Figure 5.4 therefore corresponds to the following system of equations:

$$\begin{aligned}
C &\coloneqq f_C(\epsilon_C) \\
D &\coloneqq f_D(C, \epsilon_D) \\
M_1 &\coloneqq f_{M_1}(C, D, \epsilon_{M_1}) \\
M_2 &\coloneqq f_{M_2}(C, D, M_1, \epsilon_{M_2}) \\
Y &\coloneqq f_Y(C, D, M_1, M_2, \epsilon_Y),
\end{aligned}$$

where the disturbance terms, denoted by $\{\epsilon_C, \epsilon_D, \epsilon_{M_1}\epsilon_{M_2}, \epsilon_Y\}$, are assumed to be mutually independent but otherwise arbitrarily distributed. The assumption of mutually independent disturbances implies that unobserved variables must not confound any of the exposure-mediator, exposure-outcome, mediator-mediator, or mediator-outcome relationships. Formally, the PSEs defined previously can be nonparametrically identified under the following assumptions: (f.i) conditional independence among the exposure and all the potential outcomes, (f.ii) conditional positivity of the exposure, and (f.iii) consistency of the observed and potential outcomes.

Assumption (f.i). The first of these assumptions requires that the exposure and all the potential outcomes, including potential values of the mediators, are conditionally independent given their antecedents. More specifically, this assumption requires that, for any values of $d, d^*, d^{**}, m_1, m_1^*$, and m_2, the following independence restrictions must hold: $\{M_1(d^*), M_2(d^{**}, m_1), Y(d, m_1, m_2)\} \perp\!\!\!\perp D|C$; $\{M_2(d^{**}, m_1), Y(d, m_1, m_2)\} \perp\!\!\!\perp M_1(d^*)|C, D$; and $Y(d, m_1, m_2) \perp\!\!\!\perp M_2(d^{**}, m_1^*)|C, D, M_1$. Taken together, these restrictions require that there must not be any unobserved confounding for the exposure-mediator, exposure-outcome, mediator-mediator, or mediator-outcome relationships. They also require that there must not be any exposure-induced confounders for the mediator-mediator or mediator-outcome relationships, whether these variables are observed or not. Thus, to avoid the potential for bias, we recommend that all exposure-induced confounders be included as additional mediators in analyses of PSEs. This assumption would be satisfied in data generated from a process resembling the model in Figure 5.4, where there are no unobserved or exposure-induced variables that confound the effects of the exposure or mediators.

Assumption (f.ii). Nonparametric identification of PSEs also requires a positive probability for all values of the exposure given each possible combination of the baseline confounders and the mediators.

Formally, this assumption requires that $P\left(d|c\right) > 0$ whenever $P\left(c\right) > 0$, $P\left(d|c, m_1\right) > 0$ whenever $P(c, m_1) > 0$, and $P\left(d|c, m_1, m_2\right) > 0$ whenever $P\left(c, m_1, m_2\right) > 0$. Substantively, it requires that there must be at least some chance that individuals experience all levels of the exposure within every subpopulation defined by the baseline confounders and the mediators.

Assumption (f.iii). Finally, nonparametric identification of PSEs requires that the observed values for mediators and outcome are consistent with their potential values. This assumption can be formally expressed as $M_1 = M_1(D)$, $M_2 = M_2(D, M_1)$, and $Y = Y(D, M_1, M_2)$. Substantively, it implies that there must not be multiple versions of the exposure or any mediator with heterogeneous effects on a downstream variable. It also requires that there must not be any interference between individuals, meaning that exposure and mediators experienced by one individual do not affect others.

Nonparametric identification formula. Let $\psi_{d_1,d_2,d} = \mathbb{E}\left[Y\left(d, M_1\left(d_1\right), M_2\left(d_2, M_1\left(d_1\right)\right)\right)\right]$, such that the decomposition of the total effect into PSEs from Equation 5.10 can be expressed as follows:

$$\mathbb{E}\left[Y\left(d\right) - Y\left(d^*\right)\right] = \underbrace{\psi_{d^*,d^*,d} - \psi_{d^*,d^*,d^*}}_{PSE_{D\to Y}(d,d^*)} + \underbrace{\psi_{d^*,d,d} - \psi_{d^*,d^*,d}}_{PSE_{D\to M_2\to Y}(d,d^*)} + \underbrace{\psi_{d,d,d} - \psi_{d^*,d,d}}_{PSE_{D\to M_1\rightsquigarrow Y}(d,d^*)}. \tag{5.13}$$

To identify the PSEs, it suffices to identify $\psi_{d_1,d_2,d}$ for any combination of values for d_1, d_2, and d. Under assumptions (f.i) to (f.iii), $\psi_{d_1,d_2,d}$ can be nonparametrically identified with the following expression (Daniel et al. 2015):

$$\psi_{d_1,d_2,d} = \sum_{m_2,m_1,c} \mathbb{E}\left[Y|c, d, m_1, m_2\right] P\left(m_2|c, d_2, m_1\right) P\left(m_1|c, d_1\right) P\left(c\right), \tag{5.14}$$

which is known as the *generalized mediation functional* (GMF; Zhou 2022). In this expression, $\mathbb{E}\left[Y|c, d, m_1, m_2\right]$ denotes the conditional expected value of the observed outcome Y among individuals for whom $C = c$, $D = d$, $M_1 = m_1$, and $M_2 = m_2$. In addition, $P\left(m_2|c, d_2, m_1\right)$ is the probability of experiencing level m_2 of the second mediator among those who experienced level d_2 of the exposure and level m_1 of the first mediator, and for whom $C = c$. Similarly, $P\left(m_1|c, d_1\right)$ is the probability of experiencing level m_1 of the first mediator among those who experienced level d_1 of the exposure and for whom $C = c$. As before, $P\left(c\right)$ denotes the marginal probability that $C = c$. Essentially, the GMF computes the marginal mean of the nested potential outcomes, $\psi_{d_1,d_2,d}$, by successively averaging the conditional mean of the outcome given the confounders, exposure, and mediators over the distributions of each mediator, conditional on different levels of the exposure and confounders, and then over the unconditional distribution of the confounders.

5.5 Nonparametric Estimation

In this section, we shift our attention from nonparametric identification with full population data to nonparametric estimation of the GMF and associated PSEs using a random sample. In practice, nonparametric estimation can be challenging or even impossible to implement due to the twin problems of data sparsity and the curse of dimensionality. To circumvent these challenges, we use a contrived example with only a few simplified measures. Nevertheless, nonparametric estimation remains relevant because it provides consistent estimates under weaker assumptions than parametric approaches, whenever it can be practically implemented with available data.

Nonparametric estimation of PSEs just involves substituting population quantities with their sample analogues in Equation 5.14. To illustrate, consider again the NLSY data presented in Table 5.4, which is a reproduction of Table 4.1 from the previous chapter. This table summarizes case counts and sample means, separately by maternal education C, college attendance D, unemployment status M_1, and household income M_2. Specifically, $\bar{Y}_{c,d,m_1m_2}$ denotes the sample mean of the standardized CES-D scores among respondents for whom $C = c$, $D = d$, $M_1 = m_1$, and $M_2 = m_2$, while n_{c,d,m_1m_2} represents the number of respondents with these values on each of the covariates. To simplify our analysis and enable nonparametric estimation with the NLSY, the second mediator–household income–has been recast as a binary variable, indicating whether a respondent earned over \$50,000 annually from age 35 to 39.

Based on these data, a nonparametric estimator for $\psi_{d_1,d_2,d}$ can be expressed as follows:

$$\hat{\psi}^{np}_{d_1,d_2,d} = \sum_{m_2,m_1,c} \bar{Y}_{c,d,m_1,m_2}\hat{\pi}_{m_2|c,d_2,m_1}\hat{\pi}_{m_1|c,d_1}\hat{\pi}_c \tag{5.15}$$

In this expression, $\hat{\pi}_c = \sum_{m_2,m_1,d} n_{c,d,m_1,m_2}/n$ denotes the proportion of sample members for whom $C = c$. Similarly, $\hat{\pi}_{m_1|c,d_1} = \sum_{m_2} n_{c,d_1,m_1,m_2}/\sum_{m_2,m_1} n_{c,d_1,m_1,m_2}$ denotes the proportion of sample members for whom $M_1 = m_1$ among those with $C = c$ and $D = d_1$, and $\hat{\pi}_{m_2|c,d_2,m_1} = n_{c,d_2,m_1,m_2}/\sum_{m_2} n_{c,d_2,m_1,m_2}$ denotes the proportion of sample members for whom $M_2 = m_2$ among those with $C = c$, $D = d_2$, and $M_1 = m_1$. As in previous chapters, we use "hats" to distinguish estimators from estimands and superscripts to differentiate among distinct estimators. The "np" superscript in Equation 5.15 is intended to signify that it is a nonparametric estimator for the GMF.

Based on Equation 5.15 for the GMF, nonparametric estimators for the total effect and the PSEs of interest can then be expressed as follows:

$$\begin{aligned}
\widehat{ATE}\left(d,d^*\right)^{np} &= \sum_{m_2,m_1,c} \left(\bar{Y}_{c,d,m_1,m_2}\hat{\pi}_{m_2|c,d,m_1}\hat{\pi}_{m_1|c,d} - \bar{Y}_{c,d^*,m_1,m_2}\hat{\pi}_{m_2|c,d^*,m_1}\hat{\pi}_{m_1|c,d^*}\right)\hat{\pi}_c \\
\widehat{PSE}_{D\to Y}\left(d,d^*\right)^{np} &= \sum_{m_2,m_1,c} \left(\bar{Y}_{c,d,m_1,m_2} - \bar{Y}_{c,d^*,m_1,m_2}\right)\hat{\pi}_{m_2|c,d^*,m_1}\hat{\pi}_{m_1|c,d^*}\hat{\pi}_c \\
\widehat{PSE}_{D\to M_2\to Y}\left(d,d^*\right)^{np} &= \sum_{m_2,m_1,c} \left(\hat{\pi}_{m_2|c,d,m_1} - \hat{\pi}_{m_2|c,d^*,m_1}\right)\bar{Y}_{c,d,m_1,m_2}\hat{\pi}_{m_1|c,d^*}\hat{\pi}_c \\
\widehat{PSE}_{D\to M_1\rightsquigarrow Y}\left(d,d^*\right)^{np} &= \sum_{m_2,m_1,c} \left(\hat{\pi}_{m_1|c,d} - \hat{\pi}_{m_1|c,d^*}\right)\bar{Y}_{c,d,m_1,m_2}\hat{\pi}_{m_2|c,d,m_1}\hat{\pi}_c.
\end{aligned} \tag{5.16}$$

Applying these expressions to the data in Table 5.4, we obtain nonparametric estimates of the total and path-specific effects of college attendance on depression. As in Section 3.4, our total effect estimate is -0.24, suggesting that attending college reduces CES-D scores by about one-quarter of a standard deviation, on average. The estimated direct effect, $\widehat{PSE}_{D\to Y}\left(1,0\right)^{np}$, is $-.13$, which suggests that slightly more than half of the total effect is not mediated by either unemployment or household income. The estimated path-specific effect for household income only, $\widehat{PSE}_{D\to M_2\to Y}\left(1,0\right)^{np}$, is $-.09$. This suggests that more than one-third of the total effect operates through household income but not unemployment. By contrast, the estimated path-specific effect for unemployment, $\widehat{PSE}_{D\to M_1\rightsquigarrow Y}\left(1,0\right)^{np}$, is only about $-.02$, indicating that unemployment at midlife plays a minimal role in transmitting the effect of college attendance on depression.

These estimators are only consistent if the assumptions required to nonparametrically identify the GMF and the associated PSEs are satisfied. Otherwise, they are biased and inconsistent. In our simplified illustration with the NLSY, assumption (f.i) is likely violated because we only adjusted for a single covariate–

Table 5.4: Case Counts and Sample Means for CES-D Scores (Y) by College Attendance (D), Unemployment Status (M_1), Household Income (M_2), and Maternal Education (C), NLSY.

Respondent Education	Unemployment Status	Household Income	Maternal Education			
			No College		Attended College	
			$(C = 0)$		$(C = 1)$	
			$\bar{Y}_{c,d,m_1,m_2}$	n_{c,d,m_1,m_2}	$\bar{Y}_{c,d,m_1,m_2}$	n_{c,d,m_1,m_2}
No College	Never Unemployed	$<$ \$50K ($M_2 = 0$)	.07	1219	.05	98
$(D = 0)$	$(M_1 = 0)$	$\geq$ \$50K ($M_2 = 1$)	−.17	528	−.11	74
	Ever Unemployed	$<$ \$50K ($M_2 = 0$)	.35	460	.39	36
	$(M_1 = 1)$	$\geq$ \$50K ($M_2 = 1$)	.04	60	.34	3
Attended College	Never Unemployed	$<$ \$50K ($M_2 = 0$)	−.07	211	.03	85
$(D = 1)$	$(M_1 = 0)$	$\geq$ \$50K ($M_2 = 1$)	−.30	358	−.22	247
	Ever Unemployed	$<$ \$50K ($M_2 = 0$)	.24	56	.14	18
	$(M_1 = 1)$	$\geq$ \$50K ($M_2 = 1$)	−.16	36	−.23	24

Note: CES-D scores have been standardized to have zero mean and unit variance. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch5/table_5-4`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

maternal education–at baseline, when a multitude of other factors may jointly influence the exposure, mediators, and outcome, resulting in bias due to unobserved (i.e., uncontrolled) confounding.

Furthermore, nonparametric estimation is typically complicated by sparse data and the curse of dimensionality, as discussed at length in Chapter 3. These challenges stem from our reliance on finite samples that contain many variables or include measures with many values, and they are frequently severe enough to preclude nonparametric estimation entirely. Indeed, it would not have been possible to nonparametrically estimate PSEs with the NLSY had we not artificially simplified our analysis to reduce the dimension of the data and thereby mitigate the problem of sparsity.

Thus, despite its theoretical promise of providing consistent results under weaker assumptions than other methods, nonparametric estimation may be impractical or infeasible in many social science applications. In these instances, estimation strategies that rely on parametric models offer significant advantages, albeit at the cost of more stringent assumptions about the probability distribution from which the data were sampled. In the next section, we introduce a series of parametric estimators for PSEs, which are more broadly applicable in practice.

5.6 Parametric Estimation

In this section, we discuss several parametric methods for estimating PSEs. As noted in Section 5.3.2, the $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$ is simply the natural indirect effect with respect to M_1, while the sum of the $PSE_{D \to Y}(d, d^*)$ and $PSE_{D \to M_2 \to Y}(d, d^*)$ corresponds to the natural direct effect with respect to M_1. Moreover, the $PSE_{D \to Y}(d, d^*)$ is equivalent to the $NDE_{\mathbf{M}}(d, d^*)$, that is, the multivariate natural direct effect with respect to both mediators considered together. Thus, the PSEs defined in Equation 5.10 can also be expressed as follows:

$$\begin{aligned} PSE_{D \to Y}(d, d^*) &= NDE_{\mathbf{M}}(d, d^*) \\ PSE_{D \to M_2 \to Y}(d, d^*) &= NDE_{M_1}(d, d^*) - NDE_{\mathbf{M}}(d, d^*) \\ PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*) &= NIE_{M_1}(d, d^*), \end{aligned} \tag{5.17}$$

where $NDE_{\mathbf{M}}(d, d^*)$ denotes the multivariate natural direct effect with respect to $\mathbf{M} = (M_1, M_2)$. To estimate the PSEs, then, we need only estimate the natural direct and indirect effects on the right-hand sides of these equations. In other words, we can use the methods from Chapter 3 to obtain estimates of $NIE_{M_1}(d, d^*)$ and $NDE_{M_1}(d, d^*)$, and the methods introduced in Section 5.3 of the present chapter to obtain an estimate of $NDE_{\mathbf{M}}(d, d^*)$. Then, we can simply plug these estimates into the expressions provided above to obtain estimates of the $PSE_{D \to Y}(d, d^*)$, $PSE_{D \to M_2 \to Y}(d, d^*)$, and $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$. Confidence intervals and p-values for hypothesis tests can be constructed using the nonparametric bootstrap.

In the sections that follow, we first illustrate this approach with estimators based on linear models and inverse probability weighting. We then introduce a regression-imputation estimator that is particularly useful in analyses of multiple mediators because it does not require modeling these variables (Zhou and Yamamoto 2022).

5.6.1 Estimation with Linear Models

To estimate the $PSE_{D \to Y}(d, d^*)$, $PSE_{D \to M_2 \to Y}(d, d^*)$, and $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$ with linear models, it suffices to obtain the corresponding estimates of the natural direct and indirect effects shown in Equation

5.17. These estimates can be constructed, for example, by fitting the following set of models:

$$\begin{aligned}
\mathbb{E}\left[M_1|c,d\right] &= \beta_{01} + \beta_{11}^T c + \beta_{21} d, && (5.18)\\
\mathbb{E}\left[M_2|c,d\right] &= \beta_{02} + \beta_{12}^T c + \beta_{22} d, && (5.19)\\
\mathbb{E}\left[Y|c,d,m_1\right] &= \delta_0 + \delta_1^T c + \delta_2 d + \delta_3 m_1, && (5.20)\\
\mathbb{E}\left[Y|c,d,m_1,m_2\right] &= \gamma_0 + \gamma_1^T c + \gamma_2 d + \gamma_{31} m_1 + \gamma_{32} m_2, && (5.21)
\end{aligned}$$

where $\mathbb{E}\left[M_k|c,d\right]$ for $k = 1, 2$ denotes the conditional mean of M_k given the baseline confounders and the exposure, $\mathbb{E}\left[Y|c,d,m_1\right]$ denotes the conditional mean of the outcome given the confounders, exposure, and the first mediator, and $\mathbb{E}\left[Y|c,d,m_1,m_2\right]$ denotes the conditional mean of the outcome given the confounders, exposure, and both mediators.

Under assumptions (f.i) to (f.iii) and under the additional assumption of correct model specification, consistent estimators for $NDE_{M_1}(d,d^*)$, $NIE_{M_1}(d,d^*)$, and $NDE_{\mathbf{M}}(d,d^*)$ can be obtained by fitting Equations 5.18 to 5.21 using the method of OLS and then substituting their coefficient estimates into appropriate parametric expressions. Specifically, these estimators can be constructed as follows:

$$\begin{aligned}
\widehat{NDE}_{M_1}(d,d^*)^{lm} &= \hat{\delta}_2(d-d^*)\\
\widehat{NIE}_{M_1}(d,d^*)^{lm} &= \hat{\beta}_{21}\hat{\delta}_3(d-d^*)\\
\widehat{NDE}_{\mathbf{M}}(d,d^*)^{lm} &= \hat{\gamma}_2(d-d^*), && (5.22)
\end{aligned}$$

where the "hats" distinguish estimators from estimands, as before, and where the "lm" superscript indicates that these estimators are based on linear and additive models fit by least squares. By extension, consistent estimators for the total and path-specific effects of interest can be constructed with the following expressions:

$$\begin{aligned}
\widehat{ATE}(d,d^*)^{lm} &= \widehat{NDE}_{M_1}(d,d^*)^{lm} + \widehat{NIE}_{M_1}(d,d^*)^{lm}\\
\widehat{PSE}_{D\to Y}(d,d^*)^{lm} &= \widehat{NDE}_{\mathbf{M}}(d,d^*)^{lm}\\
\widehat{PSE}_{D\to M_2\to Y}(d,d^*)^{lm} &= \widehat{NDE}_{M_1}(d,d^*)^{lm} - \widehat{NDE}_{\mathbf{M}}(d,d^*)^{lm}\\
\widehat{PSE}_{D\to M_1\rightsquigarrow Y}(d,d^*)^{lm} &= \widehat{NIE}_{M_1}(d,d^*)^{lm}. && (5.23)
\end{aligned}$$

These estimators are based on highly restrictive models, and thus they are prone to misspecification bias. In particular, the models on which these estimators are based do not allow the effects of the exposure and mediators to interact or to vary across levels of the baseline confounders. As in analyses with a single mediator, estimation with linear models can be easily adapted to accommodate interaction effects between the exposure and mediators and/or effect moderation by baseline confounders. Specifically, we can modify Equations 5.20 to 5.21 with more flexible specifications, including exposure-mediator interactions and interactions with the baseline confounders, after centering them around their sample means. These modified regressions can be used to obtain another set of estimates for the natural direct and indirect effects that compose the PSEs of interest, denoted as $\widehat{NDE}_{M_1}(d,d^*)^{lmi}$, $\widehat{NIE}_{M_1}(d,d^*)^{lmi}$, and $\widehat{NDE}_{\mathbf{M}}(d,d^*)^{lmi}$ to indicate they are based on models with interactions. These effects can then be substituted in Equation 5.17 above to obtain a corresponding set of estimates for the PSEs.

Applying these methods to data from the NLSY, the first two columns of Table 5.5 report estimates for the total and path-specific effects of college attendance on depression. Consistent with prior results, estimates for the total effect are about $-.07$, implying that attending college reduces CES-D scores by about

Table 5.5: Total and Path-Specific Effects of College Attendance on CES-D Scores as Estimated from Linear Models and Inverse Probability Weighting with the NLSY.

Estimand	Point Estimates and 95% CIs		
	Additive Linear Model (LinMod)	LinMod with $D \times M$ Interactions	Inverse Probability Weighting
$ATE\,(1,0)$	$-.070$ $[-.152, .010]$	$-.074$ $[-.158, .007]$	$-.167$ $[-.265, -.054]$
$PSE_{D\to Y}\,(1,0)$	$-.023$ $[-.105, .058]$	$-.037$ $[-.135, .061]$	$-.100$ $[-.232, .052]$
$PSE_{D\to M_2\to Y}\,(1,0)$	$-.040$ $[-.057, -.026]$	$-.033$ $[-.076, .006]$	$-.059$ $[-.138, -.003]$
$PSE_{D\to M_1\rightsquigarrow Y}\,(1,0)$	$-.006$ $[-.016, .002]$	$-.004$ $[-.014, .002]$	$-.009$ $[-.027, .004]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals, which were computed using the nonparametric bootstrap with $B = 2000$ replications. D denotes college attendance, M_1 denotes unemployment, M_2 denotes household income, and Y denotes standardized CES-D scores. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch5/table_5-5`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

0.07 standard deviations, on average. Moreover, estimates of the PSEs suggest that a substantial portion of the total effect may operate through household income only, that is, along the $D \to M_2 \to Y$ path. According to our estimates based on linear models with exposure-mediator interactions, for example, the estimated $PSE_{D\to M_2\to Y}\,(1,0)$ is about $-.03$, or roughly 40% of the total effect. By contrast, unemployment appears to play a smaller mediating role, as the estimated $PSE_{D\to M_1\rightsquigarrow Y}\,(1,0)$ is consistently close to zero. Many of our estimates, however, have wide confidence intervals, which precludes definitive conclusions about the size of the different effects. The code and data used to conduct this analysis can be accessed via links in the table footnote.

5.6.2 Estimation with Inverse Probability Weights

As with estimators based on linear models, weighting estimators for PSEs can be obtained from weighting estimators for the corresponding natural direct and indirect effects in Equation 5.17. This is accomplished by applying the procedure described in Section 3.5.3 from Chapter 3 twice, first with M_1 as the focal mediator and then with $\mathbf{M} = (M_1, M_2)$. Specifically, weighting estimators for the natural direct and indirect effects with respect to M_1 are given by the following expressions:

$$\begin{aligned}\widehat{NDE}_{M_1}\,(d,d^*)^{ipw} &= \frac{\sum I\,(D=d)\,\hat{sw}_3 Y}{\sum I\,(D=d)\,\hat{sw}_3} - \frac{\sum I\,(D=d^*)\,\hat{sw}_1 Y}{\sum I\,(D=d^*)\,\hat{sw}_1}\\ \widehat{NIE}_{M_1}\,(d,d^*)^{ipw} &= \frac{\sum I\,(D=d)\,\hat{sw}_2 Y}{\sum I\,(D=d)\,\hat{sw}_2} - \frac{\sum I\,(D=d)\,\hat{sw}_3 Y}{\sum I\,(D=d)\,\hat{sw}_3},\end{aligned} \tag{5.24}$$

where $\hat{sw}_1$, $\hat{sw}_2$, and $\hat{sw}_3$ are stabilized weights defined as

$$\hat{sw}_1 = \frac{\hat{P}\,(d^*)}{\hat{P}\,(d^*|C)},\ \hat{sw}_2 = \frac{\hat{P}\,(d)}{\hat{P}\,(d|C)},\ \text{and}\ \hat{sw}_3 = \frac{\hat{P}\,(d^*|C,M_1)\,\hat{P}\,(d)}{\hat{P}\,(d|C,M_1)\,\hat{P}\,(d^*|C)}. \tag{5.25}$$

Similarly, a weighting estimator for the multivariate natural direct effect with respect to both mediators together is given by the following expression:

$$\widehat{NDE}_{\mathbf{M}}\left(d,d^{*}\right)^{ipw}=\frac{\sum I\left(D=d\right)\hat{sw}_{4}Y}{\sum I\left(D=d\right)\hat{sw}_{4}}-\frac{\sum I\left(D=d^{*}\right)\hat{sw}_{1}Y}{\sum I\left(D=d^{*}\right)\hat{sw}_{1}}, \tag{5.26}$$

where $\hat{sw}_4$ is a stabilized weight defined as

$$\hat{sw}_{4}=\frac{\hat{P}\left(d^{*}|C,M_{1},M_{2}\right)\hat{P}\left(d\right)}{\hat{P}\left(d|C,M_{1},M_{2}\right)\hat{P}\left(d^{*}|C\right)}. \tag{5.27}$$

Then, with these estimators for natural direct and indirect effects, weighting estimators for the total and path-specific effects of interest can be constructed as follows:

$$\begin{aligned}\widehat{ATE}\left(d,d^{*}\right)^{ipw}&=\widehat{NDE}_{M_{1}}\left(d,d^{*}\right)^{ipw}+\widehat{NIE}_{M_{1}}\left(d,d^{*}\right)^{ipw}\\ \widehat{PSE}_{D\to Y}\left(d,d^{*}\right)^{ipw}&=\widehat{NDE}_{\mathbf{M}}\left(d,d^{*}\right)^{ipw}\\ \widehat{PSE}_{D\to M_{2}\to Y}\left(d,d^{*}\right)^{ipw}&=\widehat{NDE}_{M_{1}}\left(d,d^{*}\right)^{ipw}-\widehat{NDE}_{\mathbf{M}}\left(d,d^{*}\right)^{ipw},\\ \widehat{PSE}_{D\to M_{1}\rightsquigarrow Y}\left(d,d^{*}\right)^{ipw}&=\widehat{NIE}_{M_{1}}\left(d,d^{*}\right)^{ipw}.\end{aligned} \tag{5.28}$$

To summarize, estimates for PSEs based on inverse probability weighting can be constructed through the following steps:

1. **Fit models for the exposure.** That is, fit three GLMs for the exposure. The first model should include only the baseline confounders as predictors. The second should include both the baseline confounders and the first mediator M_1 as predictors, while the third should include the baseline confounders and both mediators, M_1 and M_2.

2. **Compute predicted probabilities of exposure.** For each sample member, compute the predicted probabilities $\hat{P}\left(d^{*}|C\right)$ and $\hat{P}\left(d|C\right)$ from the first GLM, the predicted probabilities $\hat{P}\left(d^{*}|C,M_{1}\right)$ and $\hat{P}\left(d|C,M_{1}\right)$ from the second GLM, and the predicted probabilities $\hat{P}\left(d^{*}|C,M_{1},M_{2}\right)$ and $\hat{P}\left(d^{*}|C,M_{1},M_{2}\right)$ from the third GLM. In addition, estimate the marginal probabilities of exposure using the sample proportions of individuals exposed to $D=d^{*}$ and $D=d$, respectively, which are denoted as $\hat{P}\left(d^{*}\right)$ and $\hat{P}\left(d\right)$.

3. **Construct inverse probability weights.** Substitute these predicted probabilities into Equations 5.25 and 5.27 to obtain the stabilized weights $\hat{sw}_1$, $\hat{sw}_2$, $\hat{sw}_3$, and $\hat{sw}_4$.

4. **Compute effect estimates.** Substitute the stabilized weights into Equations 5.24, and 5.26 to obtain estimates for $NDE_{M_1}\left(d,d^{*}\right)$, $NIE_{M_1}\left(d,d^{*}\right)$, and $NDE_{\mathbf{M}}\left(d,d^{*}\right)$. Then, substitute these estimates into Equation 5.28 to obtain estimates for the total and path-specific effects.

This procedure provides consistent estimates for the total and path-specific effects of interest provided that the assumptions needed to identify these effects are all satisfied and the models used to construct the weights are correctly specified.

Unlike estimators based on linear regressions, which involve modeling both the mediators and the outcome, the weighting approach only requires a set of models for the exposure. In applications where researchers have better knowledge of the generative process for the exposure than for the mediators or the outcome, inverse probability weighting may therefore be a preferable strategy for estimating total and path-specific effects. However, weighting estimators are particularly sensitive to model misspecification (Kang and Schafer 2007). Specifically, their estimates of causal effects tend to be more unstable if the requisite models are incorrectly specified. And even with correct models for the exposure, weighting estimators may also be more imprecise and more susceptible to finite sample bias. Using stabilized weights can partly mitigate these challenges, and the performance of weighting estimators can also sometimes be improved by censoring the weights, as discussed in Chapter 3.

The last column of Table 5.5 reports weighting estimates of the total and path-specific effects of college attendance on CES-D scores from the NLSY. We use logit models for college attendance to estimate the stabilized weights, which we also censor at the 1st and 99th percentiles to improve precision. Inverse probability weighting yields a larger estimate for the total effect compared with those based on linear models, but our findings regarding the mediating role of unemployment and household income are similar. Consistent with estimates based on linear models, the weighting estimates provide little evidence that unemployment mediates the effect of college attendance on depression at midlife, as the $\widehat{PSE}_{D \to M_1 \rightsquigarrow Y}\left(d, d^*\right)^{ipw}$ is close to zero. By contrast, the estimated $\widehat{PSE}_{D \to M_2 \to Y}\left(d, d^*\right)^{ipw}$ is $-.059$, suggesting that more than one-third of the total effect may operate through household income alone.

To conclude, inverse probability weighting offers another approach for estimating PSEs that only requires models for the exposure. Although this approach is flexible and can be implemented with a broad class of GLMs, it is generally best suited for applications where the exposure is discrete and takes on relatively few values. In applications where the exposure has many values or is continuous, the weighting approach may perform poorly due to unreliable estimates of the probabilities or densities needed to construct the weights. When the exposure is binary, ordinal, or polytomous, and the researcher is confident in their ability to model its distribution, weighting may be the preferred option, especially in analyses with multiple mediators. Otherwise, the alternative estimators discussed elsewhere may provide more accurate results.

5.6.3 Estimation with Regression Imputation

As outlined previously, weighting estimators for path-specific effects can be inefficient, susceptible to finite sample bias, and sensitive to model misspecification. In contrast, estimators based on linear models avoid some of these challenges, but their modeling demands are also onerous in analyses of multiple mediators, because they require correct regressions for each one. Moreover, the necessity that these models must all be linear makes them less suitable for applications involving categorical mediators or outcomes.

In this section, we introduce a regression-imputation approach to estimating path-specific effects that can overcome these limitations. This method requires a series of models for the conditional mean of the outcome, which need not be linear. These models can be linear regressions, but they can also be GLMs, and they can incorporate a wide variety of interactions and nonlinear terms. In addition, the regression-imputation approach does not require modeling the mediators at all. Its modeling demands are therefore much less restrictive than the estimators we have considered previously. Moreover, this approach can also accommodate many different types of exposures, mediators, and outcomes, including both discrete and continuous measures.

Recall that the PSEs of interest involve four counterfactual means: $\mathbb{E}\left[Y\left(d\right)\right]$, $\mathbb{E}\left[Y\left(d^*\right)\right]$, $\mathbb{E}\left[Y\left(d, M_1\left(d^*\right), M_2\left(d^*, M_1\left(d^*\right)\right)\right)\right]$, and $\mathbb{E}\left[Y\left(d, M_1\left(d^*\right), M_2\left(d, M_1\left(d^*\right)\right)\right)\right]$. Thus, to estimate the PSEs, it suffices to estimate, or "impute," each of these means. Under assumptions (a.i) to (a.iii) from Chapter 3, the first two means can be expressed as follows:

$$\begin{aligned}\mathbb{E}\left[Y\left(d^*\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[Y|D=d^*,C\right]\right]\\ \mathbb{E}\left[Y\left(d\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[Y|D=d,C\right]\right].\end{aligned}$$

To estimate these quantities, we could fit a model for the outcome given the exposure and the baseline confounders. We could then use this model to predict the conditional means, $\mathbb{E}\left[Y|D=d^*,C\right]$ and $\mathbb{E}\left[Y|D=d,C\right]$, for every sample member. And finally, to estimate the outer expectations, we could just average these predicted means across sample members.

Similarly, under assumptions (f.i) to (f.iii), the other two counterfactual means, $\mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d^*,M_1\left(d^*\right)\right)\right)\right]$ and $\mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d,M_1\left(d^*\right)\right)\right)\right]$, can be expressed using iterated expectations as follows:

$$\begin{aligned}\mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d^*,M_1\left(d^*\right)\right)\right)\right] &= \mathbb{E}\Big[\mathbb{E}\left[\mathbb{E}\left[Y|C,D=d,M_1,M_2\right]|C,D=d^*\right]\Big]\\ \mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d,M_1\left(d^*\right)\right)\right)\right] &= \mathbb{E}\Big[\mathbb{E}\left[\mathbb{E}\left[Y|C,D=d,M_1\right]|C,D=d^*\right]\Big].\end{aligned} \tag{5.29}$$

To estimate these quantities, we could first fit models for $\mathbb{E}[Y|C,D,M_1,M_2]$ and $\mathbb{E}[Y|C,D,M_1]$ and use them to predict the two conditional means, $\mathbb{E}\left[Y|C,D=d,M_1,M_2\right]$ and $\mathbb{E}\left[Y|C,D=d,M_1\right]$. We could then fit another set of models for these predictions, conditional on the baseline confounders and exposure, and use these models to estimate $\mathbb{E}\left[\mathbb{E}\left[Y|C,D=d,M_1,M_2\right]|C,D=d^*\right]$ and $\mathbb{E}\left[\mathbb{E}\left[Y|C,D=d,M_1\right]|C,D=d^*\right]$ for each sample member. Lastly, to estimate the outermost expectations, we could just average these predictions across sample members.

Alternatively, these counterfactual means can also be expressed as follows:

$$\begin{aligned}\mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d^*,M_1\left(d^*\right)\right)\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[Y|C,D=d,M_1,M_2\right]\frac{P\left(D=d^*\right)}{P\left(D=d^*|C\right)}\Big|D=d^*\right]\\ \mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d,M_1\left(d^*\right)\right)\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[Y|C,D=d,M_1\right]\frac{P\left(D=d^*\right)}{P\left(D=d^*|C\right)}\Big|D=d^*\right].\end{aligned} \tag{5.30}$$

These expressions suggest we could also estimate the counterfactual means by predicting $\mathbb{E}\left[Y|C,D=d,M_1,M_2\right]$, $\mathbb{E}\left[Y|C,D=d,M_1\right]$, and the probability ratio $P(D=d^*)/P(D=d^*|C)$ for each sample member. After these predictions are obtained, the outer expectations could be estimated using their sample analogs–that is, by computing weighted averages of the predicted conditional means among sample members exposed to d^*, with weights equal to $P(D=d^*)/P(D=d^*|C)$. Note that the weight used to compute these sample averages is equivalent to a stabilized inverse probability weight.

Thus, Equations 5.29 and 5.30 suggest two different approaches to estimating PSEs. Because the first approach, based on Equation 5.29, involves only a series of regression imputations, we refer to it as the *pure imputation estimator*. The second approach, based on Equation 5.30, involves both regression imputation and inverse probability weighting, so we refer to it as the *imputation-based weighting estimator* (Zhou and Yamamoto 2022).

In practice, the pure imputation estimator is implemented through the following series of steps (Trinh et al. 2021):

1. **Fit models for the outcome**. That is, fit a model for the mean of the outcome conditional on the exposure and baseline confounders, denoted by $\mathbb{E}\left[Y|C,D\right]$. Then, fit a second model for the mean

of the outcome conditional on the exposure, baseline confounders, and both mediators, denoted by $\mathbb{E}\left[Y|C,D,M_1,M_2\right]$. Finally, fit a third model for the mean of the outcome conditional on the exposure, the baseline confounders, and the first mediator only, denoted by $\mathbb{E}\left[Y|C,D,M_1\right]$.

2. **Impute the conventional potential outcomes.** Use the fitted model for $\mathbb{E}\left[Y|C,D\right]$ to impute the mean of the potential outcomes under exposure d^* by setting $D=d^*$ for all sample members, computing predicted values $\hat{\mathbb{E}}\left[Y|C,D=d^*\right]$, and then computing the sample average of these predictions, which yields an estimate for $\mathbb{E}\left[Y\left(d^*\right)\right]$. Similarly, use the same model to estimate the mean of the potential outcomes under exposure d by setting $D=d$ for all sample members, computing predicted values $\hat{\mathbb{E}}\left[Y|C,D=d\right]$, and then computing the sample average of these predictions. This yields an estimate for $\mathbb{E}\left[Y\left(d\right)\right]$.

3. **Impute the cross-world potential outcomes.** With the fitted model for $\mathbb{E}\left[Y|C,D,M_1,M_2\right]$, set $D=d$ for all sample members and compute predicted values, denoted by $\hat{\mathbb{E}}\left[Y|C,D=d,M_1,M_2\right]$. Then, use these predicted values to impute the mean of the cross-world potential outcomes under $\left\{d,M_1\left(d^*\right),M_2\left(d^*,M_1\left(d^*\right)\right)\right\}$ by fitting a model for $\hat{\mathbb{E}}\left[Y|C,D=d,M_1,M_2\right]$ conditional on the exposure and baseline confounders, setting $D=d^*$ for all sample members, and computing another set of predicted values, denoted by $\hat{\mathbb{E}}\left[\hat{\mathbb{E}}\left[Y|C,D=d,M_1,M_2\right]\middle|C,D=d^*\right]$. The sample average of these predictions yields an estimate for $\mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d^*,M_1\left(d^*\right)\right)\right)\right]$. Next, with the fitted model for $\mathbb{E}\left[Y|C,D,M_1\right]$, set $D=d$ for all sample members and compute predicted values, denoted by $\hat{\mathbb{E}}\left[Y|C,D=d,M_1\right]$. Then, use these predicted values to impute the mean of the cross-world potential outcomes under $\left\{d,M_1\left(d^*\right),M_2\left(d,M_1\left(d^*\right)\right)\right\}$ by fitting a model for $\hat{\mathbb{E}}\left[Y|C,D=d,M_1\right]$ conditional on the exposure and baseline confounders, setting $D=d^*$ for all sample members, and computing another set of predicted values, denoted by $\hat{\mathbb{E}}\left[\hat{\mathbb{E}}\left[Y|C,D=d,M_1\right]\middle|C,D=d^*\right]$. The sample average of these predictions yields an estimate for $\mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d,M_1\left(d^*\right)\right)\right)\right]$.

4. **Compute effect estimates.** Calculate estimates for the PSEs of interest by substituting the imputed potential outcomes from steps 2 and 3 into Equation 5.10.

The imputation-based weighting estimator follows a similar implementation but additionally requires an estimate for the stabilized inverse probability weight, ${}^{P(D=d^*)}/_{P(D=d^*|C)}$. To this end, we would first estimate the numerator of weight using the proportion of sample members exposed to d^*, and then we would estimate the denominator using a GLM for the exposure with the baseline confounders as predictors. We would then repeat the procedure outlined previously after modifying step 3 as follows:

3*. With the fitted model for $\mathbb{E}\left[Y|C,D,M_1,M_2\right]$, set $D=d$ for all sample members and compute predicted values, denoted by $\hat{\mathbb{E}}\left[Y|C,D=d,M_1,M_2\right]$. Then, use these predicted values to impute the mean of the cross-world potential outcomes under $\left\{d,M_1\left(d^*\right),M_2\left(d^*,M_1\left(d^*\right)\right)\right\}$ by computing a weighted average of $\hat{\mathbb{E}}\left[Y|C,D=d,M_1,M_2\right]$ among sample members exposed to d^*, with weights equal to ${}^{\hat{P}(D=d^*)}/_{\hat{P}(D=d^*|C)}$. This weighted average yields an estimate for $\mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d^*,M_1\left(d^*\right)\right)\right)\right]$. Next, with the fitted model for $\mathbb{E}\left[Y|C,D,M_1\right]$, set $D=d$ for all sample members and compute predicted values, denoted by $\hat{\mathbb{E}}\left[Y|C,D=d,M_1\right]$. Then, use these predicted values to impute the mean of the cross-world potential outcomes under $\left\{d,M_1\left(d^*\right),M_2\left(d,M_1\left(d^*\right)\right)\right\}$ by computing a weighted average of $\hat{\mathbb{E}}[Y|C,D=d,M_1]$ among sample members exposed to d^*, also with weights equal to ${}^{\hat{P}(D=d^*)}/_{\hat{P}(D=d^*|C)}$. This weighted average yields an estimate for $\mathbb{E}\left[Y\left(d,M_1\left(d^*\right),M_2\left(d,M_1\left(d^*\right)\right)\right)\right]$.

Table 5.6: Total and Path-Specific Effects of College Attendance on CES-D Scores as Estimated using Regression Imputation with NLSY.

Estimand	Point Estimates and 95% CIs	
	Pure Imputation Estimator	Imputation-based Weighting Estimator
$ATE(1,0)$	$-.070$ $[-.150, .008]$	$-.070$ $[-.150, .008]$
$PSE_{D \to Y}(1,0)$	$-.026$ $[-.109, .052]$	$-.016$ $[-.100, .062]$
$PSE_{D \to M_2 \to Y}(1,0)$	$-.039$ $[-.053, -.025]$	$-.041$ $[-.056, -.027]$
$PSE_{D \to M_1 \rightsquigarrow Y}(1,0)$	$-.006$ $[-.015, .002]$	$-.012$ $[-.021, -.004]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals, which were computed using the nonparametric bootstrap with $B = 2000$ replications. D denotes college attendance, M_1 denotes unemployment, M_2 denotes household income, and Y denotes standardized CES-D scores. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch5/table_5-6`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

These procedures yield consistent estimates of PSEs provided that assumptions (f.i) to (f.iii) are satisfied and the models used to construct the imputations are correctly specified. Because the imputation-based weighting estimator requires a model for the conditional probability of the exposure, it is best suited for applications where this variable is binary or has only a few discrete values. For both the pure imputation and the imputation-based weighting estimator, confidence intervals and p-values can be constructed using the nonparametric bootstrap.

Table 5.6 presents regression-imputation estimates for the total and path-specific effects of college attendance on depression in the NLSY. We implemented the pure imputation estimator using linear models in steps 1 and 3. For the imputation-based weighting estimator, we used linear models in step 1, and a logit model for the exposure to compute the inverse probability weights in step 3*. For simplicity, all of these models are additive in their predictors, although much more flexible specifications including interaction and nonlinear terms can be easily accommodated.

Estimates for the total effect suggest that college attendance reduces depression at midlife by about 0.07 standard deviations. And consistent with our prior findings, the regression-imputation estimates for the PSEs also suggest that a large portion of the total effect operates through household income only–that is, through the $A \to M_2 \to Y$ path. According to the pure imputation estimates, for example, the PSE for this pathway is $-.039$, which constitutes roughly 50% of the total effect. By contrast, these results again suggest that unemployment plays a minimal role in mediating the effect of education on mental health, as the estimated PSE for the $A \to M_1 \rightsquigarrow Y$ path is only $-.006$, according to the pure imputation approach. Although some of these point estimates are still accompanied by wide confidence intervals, the consistent pattern of results given by the different approaches to estimation, as shown in Tables 5.5 and 5.6, indicates that our main findings are robust to the different modeling assumptions required of each. The code and data used to produce these results are available through links provided in the table footnote.

5.7 Generalization to $K(\geq 2)$ Causally Ordered Mediators

Although we have focused mainly on PSEs involving only two mediators, the effect decomposition outlined previously can be easily generalized to accommodate an arbitrary number of causally ordered mediators. Let $\mathbf{M} = \{M_1, M_2, \ldots, M_K\}$ denote $K(\geq 2)$ mediators, and assume that for any $i < j$, M_i precedes M_j, such that no component of M_j may causally affect any component of M_i. In addition, let $\mathbf{M}_0 = \varnothing$ denote an empty set, let $\mathbf{M}_k = \{M_1, M_2, \ldots, M_k\}$ denote the set of mediators up through M_k, and let $\mathbf{M}_k(d) = \{M_1(d), M_2(d), \ldots, M_k(d)\}$ represent the potential values of these mediators under exposure d, where $M_k(d) = M_k(d, M_1(d), M_2(d, M_1(d)), \ldots)$ by definition.

With this notation, the average total effect of D on Y can be decomposed as follows:

$$\begin{aligned}\mathbb{E}[Y(d) - Y(d^*)] &= \underbrace{\mathbb{E}[Y(d, \mathbf{M}(d^*)) - Y(d^*)]}_{D \to Y} + \sum_{k=1}^{K} \underbrace{\mathbb{E}[Y(d, \mathbf{M}_{k-1}(d^*)) - Y(d, \mathbf{M}_k(d^*))]}_{D \to M_k \rightsquigarrow Y} \\ &= PSE_{D \to Y}(d, d^*) + \sum_{k=1}^{K} PSE_{D \to M_k \rightsquigarrow Y}(d, d^*). \qquad (5.31)\end{aligned}$$

This expression consists of $K+1$ PSEs, which together capture the effects of the exposure on the outcome transmitted along the causal paths $D \to Y$ and $D \to M_k \rightsquigarrow Y$ for $k = 1, 2, \ldots, K$. Specifically, the $PSE_{D \to Y}(d, d^*)$ captures a direct effect of exposure if all mediators were set to the values they would have naturally assumed for each individual under exposure d^*, while each $PSE_{D \to M_k \rightsquigarrow Y}(d, d^*)$ measures an indirect effect of exposure operating through the mediator M_k, net of all other mediators that precede it in causal order.

Similar to the decomposition with only two mediators, the PSEs from Equation 5.31 can also be expressed as functions of multivariate natural direct and indirect effects defined in terms of $\mathbf{M}_k$:

$$\begin{aligned}PSE_{D \to Y}(d, d^*) &= NDE_{\mathbf{M}}(d, d^*) \\ PSE_{D \to M_k \rightsquigarrow Y}(d, d^*) &= NDE_{\mathbf{M}_{k-1}}(d, d^*) - NDE_{\mathbf{M}_k}(d, d^*) \\ PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*) &= NIE_{M_1}(d, d^*). \qquad (5.32)\end{aligned}$$

When $k = 1$, $\mathbf{M}_{k-1}$ is an empty set, and thus the $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$ is equivalent to the natural indirect effect of the exposure on the outcome that operates through M_1. By contrast, when $k > 1$, $\mathbf{M}_{k-1}$ consists of all mediators from M_1 to M_{k-1}. In this case, each $PSE_{D \to M_k \rightsquigarrow Y}(d, d^*)$ is not equivalent to a natural indirect effect via M_k. Instead, it captures the "independent" mediating effect of M_k above and beyond that of its antecedent mediators. For example, if $K = 3$, the $PSE_{D \to M_2 \rightsquigarrow Y}(d, d^*)$ would capture the effects transmitted along the causal paths $D \to M_2 \to Y$ and $D \to M_2 \to M_3 \to Y$ but not along the paths $D \to M_1 \to M_2 \to Y$ and $D \to M_1 \to M_2 \to M_3 \to Y$.

To identify these PSEs, we must assume that the variables $\{D, M_1, ..., M_K, Y\}$ follow a DAG that encodes a nonparametric structural equation model with mutually independent disturbances, such that there is no unobserved confounding for any of the exposure-mediator, exposure-outcome, mediator-mediator, or mediator-outcome relationships. Under these conditions, or more formally, under generalized versions of assumptions (f.i) to (f.iii), the PSEs in Equation 5.32 are all nonparametrically identified (Zhou 2022; Zhou and Yamamoto 2022).

To estimate these effects in practice, we can use the methods introduced in Section 3.5 to obtain estimates of the $NIE_{M_1}(d, d^*)$ and $NDE_{M_1}(d, d^*)$, and the methods introduced in Section 5.3 to obtain estimates for each $NDE_{\mathbf{M}_k}(d, d^*)$. We can then substitute these estimates into Equation 5.32 to obtain estimates for the PSEs of interest. As illustrated previously, this estimation strategy can be implemented using either linear models or inverse probability weighting.

In addition, PSEs through $K(\geq 2)$ mediators can also be estimated using a generalized version of the regression-imputation approach introduced in Section 5.6 above. To appreciate this, note that identifying the components of Equation 5.31 requires identifying the counterfactual means, $\mathbb{E}[Y(d^*)]$, $\mathbb{E}[Y(d)]$, and $\mathbb{E}[Y(d, \mathbf{M}_k(d^*))]$ for $k = 1, 2, ..., K$. As in the setting with only two mediators, the first two of these means can be expressed as follows under assumptions (a.i) to (a.iii) from Chapter 3:

$$\begin{aligned}\mathbb{E}[Y(d^*)] &= \mathbb{E}[\mathbb{E}[Y|D = d^*, C]]\\ \mathbb{E}[Y(d)] &= \mathbb{E}[\mathbb{E}[Y|D = d, C]].\end{aligned}$$

To estimate these quantities, as before, we need only fit a model for the outcome given the exposure and the baseline confounders, use this model to predict the conditional means, $\mathbb{E}[Y|D = d^*, C]$ and $\mathbb{E}[Y|D = d, C]$, for every sample member, and then average these predictions across sample members. Similarly, under generalized versions of assumptions (f.i) to (f.iii), the other counterfactual means, $\mathbb{E}[Y(d, \mathbf{M}_k(d^*))]$ for $k = 1, 2, ..., K$, can be expressed as follows:

$$\begin{aligned}\mathbb{E}[Y(d, \mathbf{M}_k(d^*))] &= \mathbb{E}[\mathbb{E}[\mathbb{E}[Y|C, D = d, \mathbf{M}_k]|C, D = d^*]] &(5.33)\\ &= \mathbb{E}\left[\mathbb{E}[Y|C, D = d, \mathbf{M}_k]\frac{P(D = d^*)}{P(D = d^*|C)}|D = d^*\right]. &(5.34)\end{aligned}$$

The expression in the first equality can be estimated using a pure imputation approach, while the expression in the second can be estimated using an imputation-based weighting approach.

Specifically, the pure imputation estimator is implemented as follows:

1. **Fit a model for the outcome**. That is, fit a model for the mean of the outcome conditional on the exposure and baseline confounders, denoted by $\mathbb{E}[Y|C, D]$.

2. **Impute the conventional potential outcomes.** Use the fitted model for $\mathbb{E}[Y|C, D]$ to impute the mean of the potential outcomes under exposure d^* by setting $D = d^*$ for all sample members, computing predicted values $\hat{\mathbb{E}}[Y|C, D = d^*]$, and then computing the sample average of these predictions, which yields an estimate for $\mathbb{E}[Y(d^*)]$. Similarly, use the same model to estimate the mean of the potential outcomes under exposure d by setting $D = d$ for all sample members, computing predicted values $\hat{\mathbb{E}}[Y|C, D = d]$, and then computing the sample average of these predictions, which yields an estimate for $\mathbb{E}[Y(d)]$.

3. **Impute the cross-world potential outcomes.** For $k = 1, 2, \ldots, K$,

   (a) Fit a model for the mean of the outcome conditional on the exposure, baseline confounders, and the mediators $\mathbf{M}_k$, denoted by $\mathbb{E}[Y|C, D, \mathbf{M}_k]$.

   (b) With the fitted model for $\mathbb{E}[Y|C, D, \mathbf{M}_k]$, set $D = d$ for all sample members and compute a set of predicted values given by $\hat{\mathbb{E}}[Y|C, D = d, \mathbf{M}_k]$.

   (c) Use these predicted values to impute the mean of cross-world potential outcomes under $\{d, \mathbf{M}_k(d^*)\}$ by fitting a model for $\hat{\mathbb{E}}[Y|C, D = d, \mathbf{M}_k]$ conditional on the exposure and baseline confounders, setting $D = d^*$ for all sample members, and computing another set of predicted values, denoted by $\hat{\mathbb{E}}\left[\hat{\mathbb{E}}\left[Y|C, D = d, \mathbf{M}_k\right]|C, D = d^*\right]$. The sample average of these predictions yields an estimate for $\mathbb{E}\left[Y\left(d, \mathbf{M}_k\left(d^*\right)\right)\right]$.

4. Calculate estimates for the PSEs of interest by substituting the imputed potential outcomes from steps 2 and 3 into Equation 5.31.

For the imputation-based weighting estimator, step 3(c) is simply replaced by an inverse-probability-weighted average. That is, instead of fitting another model for the predicted values, $\hat{\mathbb{E}}[Y|C, D = d, \mathbf{M}_k]$, conditional on the exposure and baseline confounders, we would estimate $\mathbb{E}\left[Y\left(d, \mathbf{M}_k\left(d^*\right)\right)\right]$ by taking a weighted average of these predicted values among sample members exposed to d^*, with weights equal to ${}^{P(D=d^*)}/{}_{P(D=d^*|C)}$.

These procedures yield consistent estimates of PSEs through $K(\geq 2)$ mediators provided that generalized versions of assumptions (f.i) to (f.iii) are satisfied and the models used to construct the imputations are correctly specified. As before, inferential statistics can be constructed using the nonparametric bootstrap.

5.8 Sensitivity Analysis

While the effect decomposition discussed in previous sections accommodates multiple mediators and allows for identification and estimation of path-specific effects pertaining to each mediator, it still relies on the assumption that there is no unobserved confounding for any of the exposure-outcome, exposure-mediator, or mediator-outcome relationships. In observational studies, where the exposure is not randomly assigned, all of these assumptions must be scrutinized. If any are violated, estimates for the total and path-specific effects will be biased. In experimental studies, where the exposure is randomly assigned, the assumption that there is no unobserved confounding for the exposure-outcome and exposure-mediator relationships is met by design, but it remains possible that the mediator-outcome relationships are confounded by unobserved factors.

In this section, we introduce a series of bias formulas for PSEs in the presence of unobserved confounding. These formulas allow us to evaluate the sensitivity of estimates to hypothetical patterns of unobserved confounding and to examine how such biases might alter our conclusions about causal mediation. As in the previous two chapters, we begin by presenting the bias formulas in a general form that imposes minimal constraints on the pattern of confounding. However, these formulas involve a large number of parameters, complicating their use in practical applications. Thus, after introducing the bias formulas in their most general form, we shift our focus to specific patterns of unobserved confounding and incorporate several simplifying assumptions in order to express the bias terms as straightforward functions of a few basic sensitivity parameters. This simplified approach is more practical as it requires specifying only a handful of parameters, but its simplicity comes at the cost of stronger assumptions about the nature of unobserved confounding. For clarity, we concentrate on applications with two causally ordered mediators throughout this section and address the more general case of $K(\geq 2)$ mediators in Appendix J.

5.8.1 Nonparametric Bias Formulas

Consider a scenario in which an unobserved variable, denoted by U, affects the exposure D, both of the mediators M_1 and M_2, and the outcome Y, as depicted in Figure 5.5. Because the unobserved variable confounds the exposure-outcome, mediator-outcome, and exposure-mediator relationships, all the conditional independence assumptions required to identify PSEs are violated, and by extension, all the estimators we

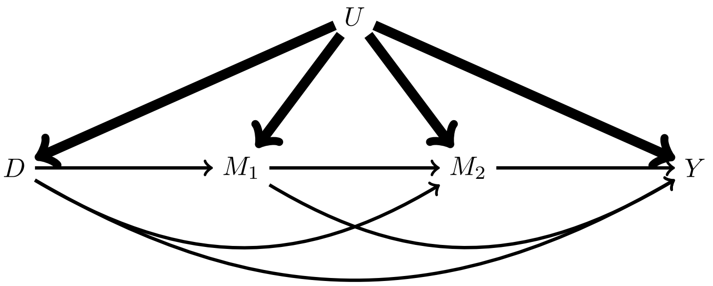


Figure 5.5: Graphical Illustration of Unobserved Exposure-outcome, Mediator-outcome, and Exposure-mediator Confounding in a Model with Two Mediators.
Note: D denotes the exposure, Y denotes the outcome, and M_j denotes mediator j. The baseline confounders C are kept implicit for visual simplicity.

have considered thus far would be biased and inconsistent. In this situation, the bias afflicting an estimator for the direct effect is given by

$$\begin{aligned}
\text{Bias}\left(\widehat{PSE}_{D\to Y}\left(d,d^*\right)\right) = & \sum_{m_1,m_2,u,c} \mathbb{E}\left[Y|c,d,m_1,m_2,u\right] P\left(m_1,m_2|c,d^*\right) P\left(c\right) \\
& \times \left(P\left(u|c,d,m_1,m_2\right) - \frac{P\left(u|c,d^*,m_1,m_2\right) P\left(u|c\right)}{P\left(u|c,d^*\right)}\right) \\
& - \sum_{m_1,m_2,u,c} \mathbb{E}\left[Y|c,d^*,m_1,m_2,u\right] P\left(m_1,m_2|c,d^*\right) P\left(c\right) \\
& \times \left(P\left(u|c,d^*,m_1,m_2\right) - \frac{P\left(u|c,d^*,m_1,m_2\right) P\left(u|c\right)}{P\left(u|c,d^*\right)}\right),
\end{aligned} \tag{5.35}$$

the bias afflicting an estimator for the path-specific effect via M_1 is given by

$$\begin{aligned}
\text{Bias}\left(\widehat{PSE}_{D\to M_1\rightsquigarrow Y}\left(d,d^*\right)\right) = & \sum_{m_1,u,c} \mathbb{E}\left[Y|c,d,m_1,u\right] P\left(m_1|c,d\right) P\left(c\right) \\
& \times \left(P\left(u|c,d,m_1\right) - \frac{P\left(u|c,d,m_1\right) P\left(u|c\right)}{P\left(u|c,d\right)}\right) \\
& - \sum_{m_1,u,c} \mathbb{E}\left[Y|c,d,m_1,u\right] P\left(m_1|c,d^*\right) P\left(c\right) \\
& \times \left(P\left(u|c,d,m_1\right) - \frac{P\left(u|c,d^*,m_1\right) P\left(u|c\right)}{P\left(u|c,d^*\right)}\right),
\end{aligned} \tag{5.36}$$

the bias afflicting an estimator for the path-specific effect via M_2 is given by

$$\begin{aligned}
\text{Bias}\left(\widehat{PSE}_{D\to M_2\to Y}\left(d,d^*\right)\right) = & \sum_{m_1,m_2,u,c} \mathbb{E}\left[Y|c,d,m_1,m_2,u\right] P\left(m_1,m_2|c,d\right) P\left(c\right) \\
& \times \left(P\left(u|c,d,m_1,m_2\right) - \frac{P\left(u|c,d,m_1,m_2\right)P\left(u|c\right)}{P\left(u|c,d\right)}\right) \\
& - \sum_{m_1,m_2,u,c} \mathbb{E}\left[Y|c,d,m_1,m_2,u\right] P\left(m_1,m_2|c,d^*\right) P\left(c\right) \\
& \times \left(P\left(u|c,d,m_1,m_2\right) - \frac{P\left(u|c,d^*,m_1,m_2\right)P\left(u|c\right)}{P\left(u|c,d^*\right)}\right) \\
& - \text{Bias}\left(\widehat{PSE}_{D\to M_1\rightsquigarrow Y}\left(d,d^*\right)\right), \qquad (5.37)
\end{aligned}$$

and the bias afflicting an estimator for the total effect is given by the sum of these three expressions.

These bias formulas are highly complex. In general, they reveal that the biases affecting estimators for PSEs depend on how the outcome differs across levels of the unobserved confounder, given the exposure, mediators, and baseline confounders. Moreover, they also reveal that the biases depend on how the unobserved confounder differs across levels of the exposure and mediators. Specifying plausible values for all the different quantities that compose these bias terms is exceptionally challenging, so it is usually necessary to introduce simplifying assumptions in order to facilitate their practical application.

5.8.2 Bias from Exposure-outcome Confounding

To this end, suppose that the unobserved variable U only confounds the relationship between the exposure D and the outcome Y. Moreover, suppose further that U is binary and that $\mathbb{E}\left[Y|c,d,m_1,m_2,U=1\right] - \mathbb{E}\left[Y|c,d,m_1,m_2,U=0\right]$ is constant across levels of C, D, M_1, and M_2, or in other words, that the difference in the mean of the outcome across levels of the unobserved confounder does not vary with the exposure, mediators, or baseline confounders. Finally, suppose that $P\left(U=1|c,d\right) - P\left(U=1|c,d^*\right)$ is constant across levels of C, or in other words, that the difference in the distribution of the unobserved confounder comparing exposure level d versus d^* does not vary with the baseline confounders.

Under these simplifying assumptions, estimators for $PSE_{D\to M_1\rightsquigarrow Y}(d,d^*)$ and $PSE_{D\to M_2\to Y}(d,d^*)$ remain unbiased. However, without adjusting for U, an estimator for the direct effect, that is, $PSE_{D\to Y}(d,d^*)$, will be subject to the following bias:

$$\text{Bias}\left(\widehat{PSE}_{D\to Y}\left(d,d^*\right)\right) = \delta_{UY|C,D,M_1,M_2} \times \delta_{DU|C}, \qquad (5.38)$$

where $\delta_{UY|C,D,M_1,M_2} = \mathbb{E}\left[Y|c,d,m_1,m_2,U=1\right] - \mathbb{E}\left[Y|c,d,m_1,m_2,U=0\right]$ and $\delta_{DU|C} = P\left(U=1|c,d\right) - P\left(U=1|c,d^*\right)$. This simple expression is a function of two sensitivity parameters, $\delta_{UY|C,D,M_1,M_2}$ and $\delta_{DU|C}$, which are easier to specify with plausible values in practice. The first sensitivity parameter $\delta_{UY|C,D,M_1,M_2}$ captures the difference in the mean of the outcome associated with a unit increase in the unobserved confounder, conditional on the baseline confounders and mediators, while the second sensitivity parameter $\delta_{DU|C}$ captures the difference in the probability of the unobserved confounder comparing level d versus d^* of the exposure, conditional on the confounders. Thus, under several simplifying assumptions, the bias in an estimator for the direct effect is equal to a "partial effect" of the unobserved confounder on the outcome multiplied by a "partial effect" of the exposure on the unobserved confounder.

5.8.3 Bias from Mediator-outcome Confounding

Consider next the scenario in which an unobserved variable U only affects the mediators M_1 and M_2 and the outcome Y, but not the exposure D. This pattern of unobserved confounding may arise in an experimental study where the exposure, but not the mediators, is randomly assigned. Random assignment of the exposure would ensure that there is no unobserved confounding of the exposure-outcome and exposure-mediator relationships. However, it would not obviate the problem of mediator-outcome confounding by unobserved factors. In this scenario, estimators for direct and path-specific effects would suffer from bias, while estimators for the total effect would remain unbiased.

Specifically, under a set of simplifying assumptions similar to those outlined previously in Section 5.8.2, the bias in an estimator for the direct effect without adjusting for U would be subject to the following bias:

$$\text{Bias}\left(\widehat{PSE}_{D\to Y}\left(d,d^{*}\right)\right)=\delta_{UY|C,D,M_1,M_2}\times\delta_{DU|C,M_1,M_2}, \tag{5.39}$$

where $\delta_{UY|C,D,M_1,M_2} = \mathbb{E}\left[Y|c,d,m_1,m_2,U=1\right] - \mathbb{E}\left[Y|c,d,m_1,m_2,U=0\right]$ and $\delta_{DU|C,M_1,M_2} = P\left(U=1|c,d,m_1,m_2\right) - P\left(U=1|c,d^{*},m_1,m_2\right)$. In this expression, $\delta_{UY|C,D,M_1,M_2}$ is defined exactly as in Section 5.8.2 above, while $\delta_{DU|C,M_1,M_2}$ represents the difference in the probability of the unobserved confounder comparing level d versus d^{*} of the exposure, now conditional on both the baseline confounders and the mediators.

Moreover, when U only confounds the mediator-outcome relationships, the biases in estimators for the PSEs pertaining to M_1 and M_2 can be expressed as follows:

$$\text{Bias}\left(\widehat{PSE}_{D\to M_1\rightsquigarrow Y}\left(d,d^{*}\right)\right)=-\delta_{UY|C,D,M_1}\times\delta_{DU|C,M_1} \tag{5.40}$$

$$\text{Bias}\left(\widehat{PSE}_{D\to M_2\to Y}\left(d,d^{*}\right)\right)=\delta_{UY|C,D,M_1}\times\delta_{DU|C,M_1}-\delta_{UY|C,D,M_1,M_2}\times\delta_{DU|C,M_1,M_2}, \tag{5.41}$$

where $\delta_{UY|C,D,M_1} = \mathbb{E}\left[Y|c,d,m_1,U=1\right] - \mathbb{E}\left[Y|c,d,m_1,U=0\right]$, $\delta_{DU|C,M_1} = P\left(U=1|c,d,m_1\right) - P\left(U=1|c,d^{*},m_1\right)$, and the sensitivity parameters $\delta_{UY|C,D,M_1,M_2}$ and $\delta_{DU|C,M_1,M_2}$ are defined exactly as before.

The form of Equations 5.39 to 5.41 demonstrates that unobserved mediator-outcome confounding distorts how our estimators partition the total effect into direct and path-specific components but does not lead to systematic error in estimates of the total effect itself. This is because the sum of these three expressions gives the bias in an estimator for the total effect that does not adjust for U, which in this case is equal to zero.

Although these bias formulas are derived under the assumption that U affects both M_1 and M_2, they are still applicable in the special case where U does not affect M_1. In this case, $\delta_{DU|C,M_1}=0$ and thus Equations 5.40 and 5.41 can be simplified to $\text{Bias}\left(\widehat{PSE}_{D\to M_1\rightsquigarrow Y}\left(d,d^{*}\right)\right)=0$ and $\text{Bias}\left(\widehat{PSE}_{D\to M_2\to Y}\left(d,d^{*}\right)\right) = -\delta_{UY|C,D,M_1,M_2}\times\delta_{DU|C,M_1,M_2}$, respectively. When U does not affect M_1, there is no unobserved confounding for the M_1-Y relationship, leading to unbiased estimates of $PSE_{D\to M_1\rightsquigarrow Y}\left(d,d^{*}\right)$.

5.8.4 Bias from Exposure-mediator Confounding

Lastly, consider the scenario in which an unobserved variable U confounds only the relationship between the exposure D and mediators M_1 and M_2. When U only confounds the exposure-mediator relationships, an estimator for the controlled direct effect with respect to the mediators M_1 and M_2 that does not adjust for U will remain unbiased. Under the assumption that the controlled direct effect does not depend on (m_1, m_2) within levels of the baseline confounders, an estimator for the direct effect that does not adjust for U, that is, $PSE_{D\to Y}(d, d^*)$, will also be unbiased in this situation. However, without adjusting for U, estimators for the PSEs pertaining to M_1 and M_2, that is, $PSE_{D\to M_1 \rightsquigarrow Y}(d, d^*)$ and $PSE_{D\to M_2\to Y}(d, d^*)$, will generally suffer from confounding bias.

If U is binary, $P(U=1|c,d) - P(U=1|c,d^*)$ is constant across levels of C, and $P(m_1, m_2|c, d, U=1) - P(m_1, m_2|c, d, U=0)$ is constant across levels of D, the biases in estimators for the PSEs pertaining to M_1 and M_2 are given by the following expressions:

$$\text{Bias}\left(\widehat{PSE}_{D\to M_1 \rightsquigarrow Y}(d, d^*)\right) = \delta_{DU|C} \times \delta_{UM_1Y|C} \tag{5.42}$$

$$\text{Bias}\left(\widehat{PSE}_{D\to M_2 \to Y}(d, d^*)\right) = \delta_{DU|C} \times \left(\delta_{UM_{12}Y|C} - \delta_{UM_1Y|C}\right) \tag{5.43}$$

In these bias formulas, $\delta_{DU|C}$ is equal to $P(U=1|c,d) - P(U=1|c,d^*)$, as in Section 5.38. This term represents the difference in the probability of the unobserved confounder comparing level d versus d^* of the exposure, conditional on the baseline confounders. The sensitivity parameter $\delta_{UM_1Y|C}$ is equal to $\sum_{m_1,c}\mathbb{E}[Y|c,d,m_1](P(m_1|c,d,U=1) - P(m_1|c,d,U=0))P(c)$. It captures how the unobserved confounder U influences the outcome Y through its effect on mediator M_1. In other words, it is similar to an “indirect effect” of the unobserved confounder on the outcome that operates through mediator M_1. Similarly, the sensitivity parameter $\delta_{UM_{12}Y|C}$ is equal to $\sum_{m_1,m_2,c}\mathbb{E}[Y|c,d,m_1,m_2](P(m_1,m_2|c,d,U=1) - P(m_1,m_2|c,d,U=0))P(c)$. It captures how the unobserved confounder U influences the outcome Y through its effect on both mediators M_1 and M_2. In other words, it is similar to an “indirect effect” of the unobserved confounder on the outcome that operates through both mediators together.

Although the bias formulas in Equations 5.42 and 5.43 are derived under the assumption that U affects both M_1 and M_2, they are still applicable in the special case where U does not affect M_2. In this situation, $\delta_{UM_{12}Y|C} = \delta_{UM_1Y|C}$, and thus Equation 5.43 can be simplified to Bias $\left(\widehat{PSE}_{D\to M_2\to Y}(d, d^*)\right) = 0$. This is because when U does not affect M_2, there is no unobserved confounding for the direct effect of D on M_2 or the effect of M_2 on Y, leading to unbiased estimates of $PSE_{D\to M_2\to Y}(d, d^*)$.

5.8.5 Bias-adjusted Effect Estimates

Table 5.7 summarizes the bias formulas for each of our target estimands under the simplifying assumptions outlined previously. With these bias formulas, a formal sensitivity analysis proceeds by reevaluating the focal effect estimates across different hypothetical patterns of unobserved confounding. We would begin this analysis by specifying the bias formulas with plausible values for their sensitivity parameters, and then we would construct a set of bias-adjusted effect estimates by subtracting the bias terms from their corresponding point estimates.

Specifically, a set of bias-adjusted estimates for the total, direct, and path-specific effects of interest can

Table 5.7: Simplified Bias Formulas for Total and Path-Specific Effects.

Estimator	Type of Confounding		
	$D \leftarrow U \rightarrow Y$	$(M_1, M_2) \leftarrow U \rightarrow Y$	$D \leftarrow U \rightarrow (M_1, M_2)$
$\widehat{PSE}_{D \rightarrow Y}(d, d^*)$	$\delta_{UY\|C,D,M_1,M_2}\delta_{DU\|C}$	$\delta_{UY\|C,D,M_1,M_2}\delta_{DU\|C,M_1,M_2}$	0
$\widehat{PSE}_{D \rightarrow M_1 \rightsquigarrow Y}(d, d^*)$	0	$-\delta_{UY\|C,D,M_1}\delta_{DU\|C,M_1}$	$\delta_{DU\|C}\delta_{UM_1Y\|C}$
$\widehat{PSE}_{D \rightarrow M_2 \rightarrow Y}(d, d^*)$	0	$\delta_{UY\|C,D,M_1}\delta_{DU\|C,M_1}$ $-\delta_{UY\|C,D,M_1,M_2}\delta_{DU\|C,M_1,M_2}$	$\delta_{DU\|C}\left(\delta_{UM_{12}Y\|C} - \delta_{UM_1Y\|C}\right)$
$\widehat{ATE}(d, d^*)$	$\delta_{UY\|C,D,M_1,M_2}\delta_{DU\|C}$	0	$\delta_{DU\|C}\delta_{UM_{12}Y\|C}$

Note: These bias formulas variously assume that U is binary; $P(U = 1|c, d, m_1, m_2) - P(U = 1|c, d^*, m_1, m_2)$ is constant across levels of C, M_1 and M_2; $\mathbb{E}[Y|c, d, m_1, m_2, U = 1] - \mathbb{E}[Y|c, d, m_1, m_2, U = 0]$ is constant across levels of C, D, M_1 and M_2; $P(m_1, m_2|c, d, U = 1) - P(m_1, m_2|c, d, U = 0)$ is constant across levels of the exposure D; and for the biases under exposure-mediator confounding specifically, that the controlled direct effect does not depend on (m_1, m_2) within levels of the baseline confounders.

be expressed as follows:

$$
\begin{aligned}
\widehat{ATE}(d, d^*)^{adj} &= \widehat{ATE}(d, d^*) - \text{Bias}\left(\widehat{ATE}(d, d^*)\right) \\
\widehat{PSE}_{D \rightarrow Y}(d, d^*)^{adj} &= \widehat{PSE}_{D \rightarrow Y}(d, d^*) - \text{Bias}\left(\widehat{PSE}_{D \rightarrow Y}(d, d^*)\right) \\
\widehat{PSE}_{D \rightarrow M_1 \rightsquigarrow Y}(d, d^*)^{adj} &= \widehat{PSE}_{D \rightarrow M_1 \rightsquigarrow Y}(d, d^*) - \text{Bias}\left(\widehat{PSE}_{D \rightarrow M_1 \rightsquigarrow Y}(d, d^*)\right) \\
\widehat{PSE}_{D \rightarrow M_2 \rightarrow Y}(d, d^*)^{adj} &= \widehat{PSE}_{D \rightarrow M_2 \rightarrow Y}(d, d^*) - \text{Bias}\left(\widehat{PSE}_{D \rightarrow M_2 \rightarrow Y}(d, d^*)\right),
\end{aligned}
$$

where $\widehat{ATE}(d, d^*)$, $\widehat{PSE}_{D \rightarrow Y}(d, d^*)$, $\widehat{PSE}_{D \rightarrow M_1 \rightsquigarrow Y}(d, d^*)$ and $\widehat{PSE}_{D \rightarrow M_2 \rightarrow Y}(d, d^*)$ each denote an estimator from Sections 5.5 to 5.6, and the "adj" superscript indicates that they have been adjusted for bias due to an assumed pattern of unobserved confounding. The degree to which our inferences are sensitive to unobserved confounding can be assessed by evaluating the bias-adjusted estimates across a range of plausible values for the sensitivity parameters and examining whether the adjusted estimates depart from hypothesized patterns. Confidence intervals for the bias-adjusted estimates can be constructed using the nonparametric bootstrap.

In practice, we could use the observed confounders to suggest plausible values for the sensitivity parameters. For example, suppose that we would like to assess the robustness of our estimated PSE via M_1 to unobserved mediator-outcome confounding. If we have an observed binary confounder Z as a component of C, we could fit a linear model of Y on C, D, and M_1, whose coefficient on Z would provide a candidate value for $\delta_{UY|C,D,M_1}$. We could also fit a linear model of Z on D, M_1, and all the other components of C, whose coefficient on D would provide a plausible value for $\delta_{DU|C,M_1}$. By combining these values for $\delta_{UY|C,D,M_1}$ and $\delta_{DU|C,M_1}$, we could assess the amount of bias that would result if an unobserved variable confounded the mediator-outcome relationships in exactly the same way as the observed confounder Z. In the next section, we illustrate this approach to specifying values for the sensitivity parameters in a mediation analysis of media framing effects.

5.9 An Empirical Illustration: The Effect of Media Framing on Immigration Attitudes

In this section, we illustrate the methods outlined previously by reanalyzing data from Brader et al. (2008; see also Imai and Yamamoto 2013 and Zhou and Wodtke 2019). This study aimed to understand the effect of negative media framing on support for immigration in the U.S. The researchers conducted a survey experiment with a nationally representative sample of 354 white non-Hispanic adults, asking respondents to read a mock news report on immigration. The news report randomly manipulated both the ethnicity of the featured immigrant and the tone of the story using a 2×2 design. Respondents were presented with a story featuring either a white European immigrant or a Latino immigrant, and the story emphasized either the benefits or the costs of immigration. After reading the story, respondents reported their beliefs about the harms of immigration, their feelings about increased immigration, and their overall level of support for immigration (Zhou and Wodtke 2019).

Brader et al. (2008) found that stories featuring a Latino immigrant (rather than a white European immigrant) and a negative frame emphasizing the costs (rather than the benefits) of immigration had a large negative effect on support for immigration. They also reported that a substantial part of this effect was mediated by anxiety about increased immigration, while beliefs about the harms of immigration did not appear to play an important mediating role. However, the researchers analyzed the mediating role of beliefs and emotions separately, assuming that respondent anxiety is not affected by perceptions of the harms associated with immigration, which seems unlikely and appears to be inconsistent with their own data (Imai and Yamamoto 2013; Zhou and Wodtke 2019).

In our analysis, we allow the two mediators to be causally related, assuming that the perceived harms of immigration causally precede respondent anxiety (Imai and Yamamoto 2013; Miller 2007). Based on this assumption, the pathways transmitting the effect of media framing can be represented by a DAG similar to the top panel of Figure 5.4. In this graph, the outcome, Y, is a measure of support for immigration on a five-point scale (standardized to have zero mean and unit variance); the exposure, D, denotes receipt of a news story featuring both a Latino immigrant and a negative frame emphasizing the costs of immigration; the first mediator, M_1, is a measure of the perceived harm of immigration on a seven-point scale; the second mediator, M_2, is the level of anxiety expressed by the respondent on a ten-point scale; and finally, the baseline covariates C include measures of gender, age, education, and income. Although exposure to negative media framing is randomly assigned, adjustment for a set of baseline covariates is necessary because the mediator-outcome relationships may still be confounded by background characteristics of the respondents.

With these data, we first computed estimates for the total and path-specific effects of negative media framing on support for immigration using linear models and inverse probability weighting. The estimates based on linear models with and without exposure-mediator interactions are shown in the first two columns of Table 5.8, while the last column reports the weighting estimates. Both approaches to estimation yield fairly similar results.

Consistent with the original study, we find a substantial total effect of negative framing. Estimates based on linear models suggest that exposure to a news story featuring both a Latino immigrant and a negative frame emphasizing the costs of immigration reduces support by about .43 standard deviations, on average. Moreover, our analysis of the PSEs suggests that over one-half of the total effect cannot be explained by either respondent anxiety or the perceived harms of immigration. According to linear models with exposure-mediator interactions, for example, the estimated $PSE_{D \to Y}(1, 0)$ is about $-.28$, representing over 60% of

Table 5.8: Total and Path-Specific Effects of Negative Media Framing on Support for Immigration as Estimated from Linear Models and Inverse Probability Weights.

Estimand	Point Estimates and 95% CIs		
	Additive Linear Model (LinMod)	LinMod with $D \times M$ Interactions	Inverse Probability Weighting
$ATE(1,0)$	$-.43$ $[-.68, -.20]$	$-.43$ $[-.68, -.20]$	$-.45$ $[-.71, -.21]$
$PSE_{D\to Y}(1,0)$	$-.23$ $[-.44, -.01]$	$-.28$ $[-.51, -.03]$	$-.31$ $[-.57, -.03]$
$PSE_{D\to M_2\to Y}(1,0)$	$-.08$ $[-.16, -.02]$	$-.05$ $[-.18, .04]$	$-.02$ $[-.16, .09]$
$PSE_{D\to M_1\rightsquigarrow Y}(1,0)$	$-.13$ $[-.27, .01]$	$-.11$ $[-.26, .01]$	$-.13$ $[-.29, .01]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals, which were computed using the nonparametric bootstrap with $B = 2000$ replications. D denotes the exposure (a news story featuring both a Latino immigrant and a negative frame emphasizing the costs of immigration), M_1 denotes perceived harm of immigration, M_2 denotes respondent anxiety, and Y denotes support for immigration (standardized to have zero mean and unit variance). The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch5/table_5-8`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/Brader_et_al2008`.

the total effect.

The $PSE_{D\to M_2\to Y}(1,0)$ in this analysis captures the independent mediating effect of respondent anxiety, above and beyond that of perceptions about the harms of immigration. According to linear models with exposure-mediator interactions, the estimated PSE via respondent anxiety is about $-.05$, or roughly 12% of the total effect. By contrast, corresponding estimates for the $PSE_{D\to M_1\rightsquigarrow Y}(1,0)$, which captures the mediating role of perceptions about the harms associated with immigration, are generally larger, although they are imprecisely estimated and their confidence intervals consistently span zero.

In addition, we estimated the same effects using the regression-imputation approach.[2] To illustrate its flexibility, we report estimates based on outcome models with no interaction terms, with exposure-mediator interactions, and with both exposure-mediator and exposure-confounder interactions. These results are presented in Table 5.8. Estimates from all three model specifications suggest that about one-quarter of the total effect of negative framing is transmitted by the perceived harms of immigration, while the independent mediating effect of respondent anxiety is relatively small. Considered altogether, then, the results from Tables 5.8 and 5.9 cast doubt on the original study's conclusion that a substantial proportion of the framing effect is mediated by anxiety about increased immigration. When perceptions about the harms of immigration are properly accounted for, the independent mediating role of emotions appears to be fairly limited.

Because our estimates suggest that most of the total effect of media framing is not mediated by perceived harms or anxiety, we conduct a sensitivity analysis for our estimates of the direct effect. We focus specifically on their sensitivity to unobserved confounding of the mediator-outcome relationships, since random assignment of the exposure in this study ensures that there is no unobserved confounding for the exposure-

[2] In experimental studies where the exposure is randomly assigned and thus independent of the baseline confounders by design, the pure imputation and imputation-based weighting estimators are equivalent because both approaches need only compute an unweighted average of the predicted values, $\hat{\mathbb{E}}[Y|C, D = d, \mathbf{M}_k]$, among sample members exposed to $D = d^*$ in step 3 of the estimation procedure.

Table 5.9: Total and Path-Specific Effects of Negative Media Framing on Support for Immigration as Estimated from the Regression-Imputation Approach.

Estimand	Point Estimates and 95% CIs		
	Outcome Models without Interactions	Outcome Models with $D \times M$ interactions	Outcome Models with $D \times \{M, C\}$ interactions
$ATE(1,0)$	$-.43\ [-.67, -.18]$	$-.43\ [-.67, -.18]$	$-.45\ [-.69, -.19]$
$PSE_{D \to Y}(1,0)$	$-.22\ [-.44, .01]$	$-.28\ [-.51, -.01]$	$-.28\ [-.51, -.02]$
$PSE_{D \to M_2 \to Y}(1,0)$	$-.08\ [-.16, -.02]$	$-.05\ [-.18, .04]$	$-.05\ [-.17, .04]$
$PSE_{D \to M_1 \rightsquigarrow Y}(1,0)$	$-.13\ [-.29, .02]$	$-.11\ [-.27, .01]$	$-.12\ [-.31, .02]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals, which were computed using the nonparametric bootstrap with $B = 2000$ replications. D denotes the exposure (a news story featuring both a Latino immigrant and a negative frame emphasizing the costs of immigration), M_1 denotes perceived harm of immigration, M_2 denotes respondent anxiety, and Y denotes support for immigration (standardized to have zero mean and unit variance). The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch5/table_5-8`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/Brader_et_al2008`.

mediator and exposure-outcome relationships. Suppose there exists a binary unobserved confounder U that affects the perceived harms of immigration (M_1), anxiety (M_2), and support for immigration (Y). In this scenario, estimates for the $PSE_{D \to Y}(1,0)$ suffer from a bias given by $\delta_{UY|C,D,M_1,M_2} \times \delta_{DU|C,M_1,M_2}$, where $\delta_{UY|C,D,M_1,M_2}$ denotes the average difference in the mean of the outcome between individuals with $U = 1$ and those with $U = 0$ conditional on the exposure, mediators, and baseline confounders, while $\delta_{DU|C,M_1,M_2}$ denotes the difference in the prevalence of U between exposed and unexposed individuals conditional on the mediators and baseline confounders.

To gain some insight into the likely sign of $\delta_{UY|C,D,M_1,M_2}$ and $\delta_{DU|C,M_1,M_2}$, suppose that U is a binary variable indicating that a respondent comes from an upper-class background. This might lead to stronger support for immigration, suggesting that $\delta_{UY|C,D,M_1,M_2} > 0$. Because the exposure is randomly assigned, the prevalence of U should be similar between sample members exposed and unexposed to negative media framing. However, because upper-class background (U) and negative framing (D) may both affect the perceived harms of immigration (M_1) and a respondent's level of anxiety (M_2), the conditional association between D and U given M_1, M_2, and C can deviate from zero. Specifically, because M_1 and M_2 are both colliders of D and U, the conditional association between the exposure and unobserved confounder could be positive, particularly if the effects of D and U on the mediators are in opposing directions. In this scenario, the bias term, $\delta_{UY|C,D,M_1,M_2} \times \delta_{DU|C,M_1,M_2}$, would be positive. Because all our estimates for the direct effect are negative, a positive bias implies that the magnitude of this effect might be even greater than already indicated in Tables 5.8 and 5.9. This line of reasoning suggests that our inferences about the direct effect of media framing are robust.

We can also use observed binary covariates to derive plausible values for the sensitivity parameters, as outlined in Section (5.8). Specifically, we consider two such variables in the data from Brader et al. (2008): whether the respondent is female, and whether the respondent holds a college degree or higher. Figure

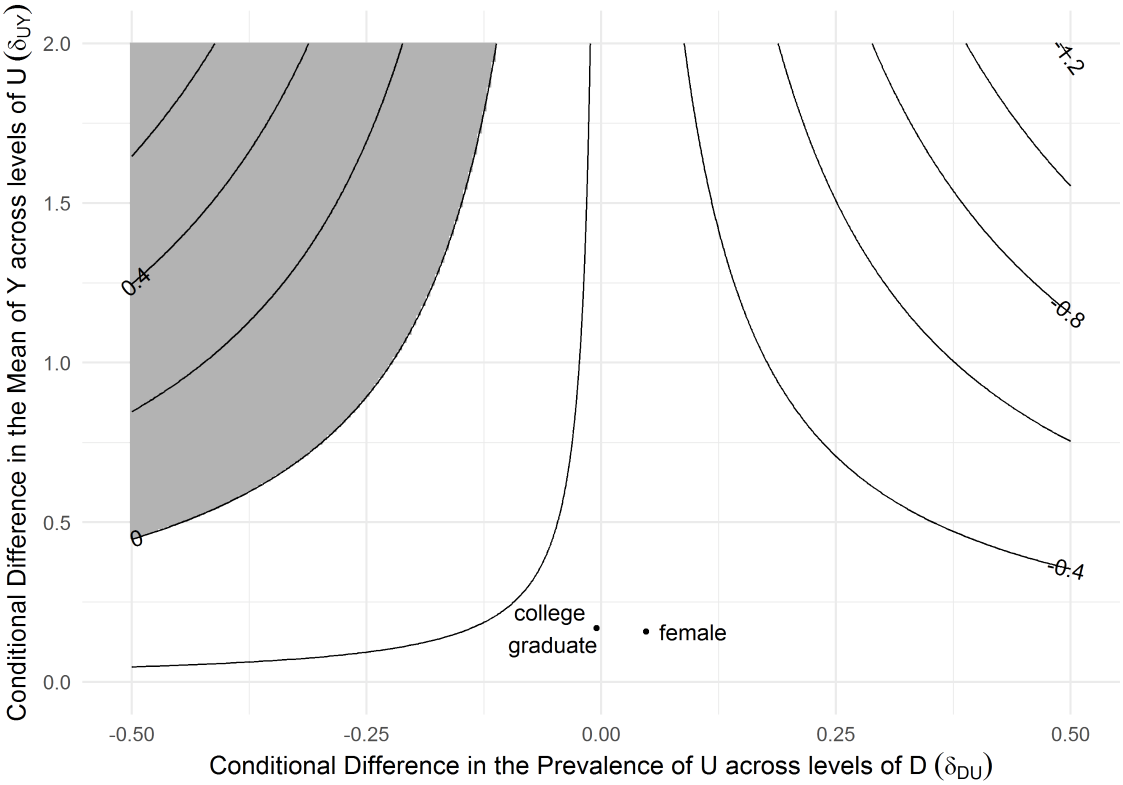


Figure 5.6: Bias-adjusted Estimates of the Direct Effect of Issue Framing on Support for Immigration. Note: The contours represent the bias-adjusted estimates of the direct effect ($PSE_{D\to Y}$) plotted as a function of $\delta_{UY|C,D,M_1,M_2}$ and $\delta_{DU|C,M_1,M_2}$. The grey area shows the values of $\delta_{UY|C,D,M_1,M_2}$ and $\delta_{DU|C,M_1,M_2}$ that would reverse the sign of the estimated $PSE_{D\to Y}$. The annotated points represent the $\delta_{UY|C,D,M_1,M_2}$ and $\delta_{DU|C,M_1,M_2}$ values that would result if the unobserved variable U confounded the mediator-outcome relationships in exactly the same way as one of the observed covariates. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch5/figure_5-6`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/Brader_et_al2008`.

5.6 shows the contours of bias-adjusted estimates for the direct effect at different values of $\delta_{UY|C,D,M_1,M_2}$ and $\delta_{DU|C,M_1,M_2}$, as well as those corresponding to an unobserved variable that mimics the confounding influence of gender and possession of a college degree. To obtain the bias-adjusted estimates, we use the point estimate for the $PSE_{D\to Y}(1,0)$ from the regression-imputation approach based on outcome models with no interaction terms. The original estimate for this effect ($-.22$) based on these models can be explained away by unobserved confounding only when $\delta_{UY|C,D,M_1,M_2}$ and $\delta_{DU|C,M_1,M_2}$ are of opposite sign and much larger than the values suggested by gender or college graduation. We conclude that our inferences are quite robust to unobserved mediator-outcome confounding. Links to the code and data used for this analysis are provided in the footnotes of the tables and figure.

5.10 Summary

In this chapter, we examined different methods for analyzing causal mediation with multiple mediators. We began with a discussion of two common approaches, which we referred to as the one-mediator-at-time approach and the multiple-mediators-as-a-whole approach. The one-mediator-at-a-time approach is appropriate only when the different mediators are causally unrelated and their relationships with each other and with the outcome are unconfounded. These conditions are strong, untestable, and unrealistic in many social science applications. The multiple-mediators-as-a-whole approach, by contrast, accommodates causally related mediators as well as several forms of mediator-mediator and mediator-outcome confounding. However, by treating all the mediators collectively, this approach cannot isolate the explanatory role of each mediator taken separately. Thus, it is most useful in applications where the mediators can be treated as different indicators of a single construct.

Given the limitations of these two approaches, we then introduced a third method that accommodates causally related mediators while still isolating their respective contributions to the total effect. This approach is based on a decomposition of the total effect into a direct effect and a series of PSEs. Each of these PSEs quantifies the net contribution of a specific mediator to the total effect, above and beyond that of its antecedent mediators. The components of this decomposition can all be nonparametrically identified if there is no unobserved confounding for any of the exposure-mediator, exposure-outcome, mediator-mediator, or mediator-outcome relationships and provided that there is no exposure-induced confounding for the mediator-mediator and mediator-outcome relationships. Compared with the assumptions required of the one-mediator-at-a-time approach, these conditions are weaker, as they permit causal relationships among the mediators.

After defining PSEs and outlining the conditions under which they can be identified, we described several estimation strategies that are based on linear models, inverse probability weighting, and regression imputation, highlighting the connections between these approaches and those discussed in prior chapters. Estimation using linear models is best suited for applications where both the outcome and the mediators are continuous, although there are situations where it may also perform reasonably well when these variables are binary, ordinal, or counts. This approach can easily accommodate exposure-mediator interactions as well as effect moderation by baseline confounders. The weighting approach requires models for the exposure rather than models for the mediators and outcome. Moreover, it can be implemented using a broad class of GLMs for the exposure. Despite this flexibility, it is generally best suited for applications where the exposure is discrete and takes on relatively few values. In applications where the exposure has many values or is continuous, inverse probability weighting tends to perform poorly due to unreliable estimates of the conditional probabilities needed to construct the weights.

Regression imputation is similar to the approach based on linear models in that it also requires modeling the outcome as a function of the exposure, baseline confounders, and various sets of mediators. However, regression imputation does not require that these models are linear, and it does not necessitate modeling the mediators at all. We introduced two variants of this approach: a pure imputation estimator that only involves fitting a set of outcome models, and an imputation-based weighting estimator that involves a combination of outcome modeling and inverse probability weighting. The pure imputation estimator is more appealing in applications with many valued or continuous exposures, as it does not require constructing inverse probability weights. The imputation-based weighting estimator, on the other hand, is particularly attractive when the assignment mechanism for the exposure is known, as in a standard randomized experiment.

To maintain our focus and simplify the presentation of complex material, this chapter omitted certain advanced topics related to analyses of multiple mediators. These include methods for effect decomposition in the presence of exposure-induced confounding and methods for analyzing repeated measures of the exposure and mediators. We refer researchers interested in these topics to the specialized literature on interventional analogues of PSEs (Lin and VanderWeele 2017; Vansteelandt and Daniel 2017) and mediation analysis with time-varying measures (VanderWeele and Tchetgen Tchetgen 2017; Zheng and van der Laan 2017).

At this point, we have covered a wide range of estimands across the preceding chapters, including natural effects, controlled direct effects, interventional effects, and path-specific effects. Table 5.10 summarizes these quantities and their key identification conditions, highlighting in particular the independence assumptions that rule out various forms of unobserved or exposure-induced confounding.

With such a variety of estimands, readers may wonder which to prioritize and what types of research questions each one addresses. In general, the average total effect ($ATE\,(d, d^*)$) addresses broad questions like, “Does the exposure affect the outcome, and by how much?” Natural direct and indirect effects, whether defined in terms of a single mediator ($NDE\,(d, d^*)$ and $NIE\,(d, d^*)$) or a multivariate set of mediators ($NDE_{\mathbf{M}}\,(d, d^*)$ and $NDE_{\mathbf{M}}\,(d, d^*)$), address more specific questions, such as, “Why does the exposure affect the outcome? Is it because the exposure influences one or more intermediate variables, which then shape the outcome in turn?” In analyses of multiple mediators, path-specific effects ($PSE_{D \to Y}\,(d, d^*)$, $PSE_{D \to M_1 \rightsquigarrow Y}\,(d, d^*)$, and $PSE_{D \to M_k \rightsquigarrow Y}\,(d, d^*)$) further unpack these types of “why” questions by revealing which intermediate variables play the most important role in linking the exposure to the outcome. Both natural and path-specific effects are largely descriptive. They merely trace the causal process through which the effect of an exposure operates via mediators. Researchers should focus on these estimands when their goal is to describe how causal chains unfold in the real world.

In contrast, the controlled direct effect ($CDE\,(d, d^*, m)$) is more prescriptive. It addresses questions like, “How would the exposure affect the outcome if we intervened to set a mediator at a specific level for everyone?” Researchers should focus on controlled direct effects if they want to understand how altering a mediator would modify the impact of an exposure on the outcome. Interventional effects ($OE\,(d, d^*)$, $IDE\,(d, d^*)$, and $IIE\,(d, d^*)$) are similarly prescriptive, in that they capture the impact of joint interventions on both the exposure and the distribution of a focal mediator. However, unlike controlled direct effects, interventional direct and indirect effects aim to emulate the operation of a causal chain from the exposure to the outcome at the population level. In this way, they occupy a middle ground between descriptive and prescriptive estimands. Researchers might focus on interventional effects because they want to explore how simultaneous shifts in both the exposure and the population distribution of a mediator would influence the outcome, or because they wish to describe a causal process operating through a focal mediator in the presence of exposure–induced confounding. While causal chains at the individual level can be exceedingly difficult to analyze when exposure-induced confounders are present, interventional effects allow researchers to study mediation under weaker assumptions–even in the presence of exposure-induced confounding–by focusing on stochastic interventions that reflect the operation of causal chains at the population level. Together, this diverse collection of estimands allows researchers to answer a broad spectrum of causal questions, and the methods detailed in the past three chapters offer many different approaches for finding defensible answers.

Table 5.10: Summary of Estimands and Key Identification Conditions.

Estimand	Definition	Key Identification Assumptions
$ATE\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d\right)-Y\left(d^*\right)\right]$	$Y\left(d,m\right)\perp D\vert C$
$NDE\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d,M\left(d^*\right)\right)-Y\left(d^*,M\left(d^*\right)\right)\right]$	$Y\left(d,m\right)\perp D\vert C$; $Y\left(d,m\right)\perp M\vert C,D=d$; $M\left(d\right)\perp D\vert C$; $Y\left(d,m\right)\perp M\left(d^*\right)\vert C$
$NIE\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d,M\left(d\right)\right)-Y\left(d,M\left(d^*\right)\right)\right]$	
$CDE\left(d,d^*,m\right)$	$\mathbb{E}\left[Y\left(d,m\right)-Y\left(d^*,m\right)\right]$	$Y\left(d,m\right)\perp D\vert C$; $Y\left(d,m\right)\perp M\vert C,D=d$
$OE\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d,\mathcal{M}\left(d\vert C\right)\right)-Y\left(d^*,\mathcal{M}\left(d^*\vert C\right)\right)\right]$	$Y\left(d,m\right)\perp D\vert C$; $Y\left(d,m\right)\perp M\vert C,D=d$; $M\left(d\right)\perp D\vert C$
$IDE\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d,\mathcal{M}\left(d^*\vert C\right)\right)-Y\left(d^*,\mathcal{M}\left(d^*\vert C\right)\right)\right]$	
$IIE\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d,\mathcal{M}\left(d\vert C\right)\right)-Y\left(d,\mathcal{M}\left(d^*\vert C\right)\right)\right]$	
$NDE_{\mathbf{M}}\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d,\mathbf{M}\left(d^*\right)\right)-Y\left(d^*,\mathbf{M}\left(d^*\right)\right)\right]$	$Y\left(d,\mathbf{m}\right)\perp D\vert C$; $Y\left(d,\mathbf{m}\right)\perp \mathbf{M}\vert C,D=d$; $\mathbf{M}\left(d\right)\perp D\vert C$; $Y\left(d,\mathbf{m}\right)\perp \mathbf{M}\left(d^*\right)\vert C$
$NIE_{\mathbf{M}}\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d,\mathbf{M}\left(d\right)\right)-Y\left(d,\mathbf{M}\left(d^*\right)\right)\right]$	
$PSE_{D\rightarrow Y}\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d,\mathbf{M}\left(d^*\right)\right)-Y\left(d^*,\mathbf{M}\left(d^*\right)\right)\right]$	$Y\left(d,\mathbf{m}_k\right)\perp D\vert C$; $Y\left(d,\mathbf{m}_k\right)\perp \mathbf{M}_k\vert C,D=d$; $\mathbf{M}_k\left(d\right)\perp D\vert C$; $Y\left(d,\mathbf{m}_k\right)\perp \mathbf{M}_k\left(d^*\right)\vert C$ for all $k=1,...,K$
$PSE_{D\rightarrow M_1\rightsquigarrow Y}\left(d,d^*\right)$	$\mathbb{E}\left[Y\left(d,M_1\left(d\right)\right)-Y\left(d,M_1\left(d^*\right)\right)\right]$	
$PSE_{D\rightarrow M_k\rightsquigarrow Y}\left(d,d^*\right)$	$NDE_{\mathbf{M}_{k-1}}\left(d,d^*\right)-NDE_{\mathbf{M}_k}\left(d,d^*\right)$	

Note: $\mathbf{M}=\left\{M_1,M_2,\ldots,M_K\right\}$ denotes K mediators in causal order, $\mathbf{M}_k=\left\{M_1,M_2,\ldots,M_k\right\}$ denotes the set of mediators up through M_k, and $\mathbf{M}_k\left(d\right)=\left\{M_1\left(d\right),M_2\left(d\right),\ldots,M_k\left(d\right)\right\}$ represents the potential values of these mediators under exposure d, where $M_k\left(d\right)=M_k\left(d,M_1\left(d\right),M_2\left(d,M_1\left(d\right)\right),\ldots\right)$; $\mathbf{M}_k\left(d^*\right)$ is defined analogously. In addition, $\mathcal{M}\left(d\vert C\right)$ denotes a random draw from the distribution of the mediator under exposure d among individuals with confounders C, and $\mathcal{M}\left(d^*\vert C\right)$ is defined analogously. For nonparametric identification, each of these estimands also requires positivity and consistency assumptions.

Chapter 6

Mediation Analysis with Robust Estimation Methods

In the previous three chapters, we introduced methods for analyzing causal mediation in applications with baseline confounding, exposure-induced confounding, and multiple mediators. For each scenario, we defined a set of total, direct, and indirect effects, outlined the assumptions required for identifying these effects, and illustrated how they can be estimated nonparametrically with low-dimensional, discrete data. However, nonparametric analysis of causal mediation is often impractical due to data sparsity and the curse of dimensionality. Thus, for all estimands defined in this book, we outlined several parametric estimators based on regression, simulation, and weighting techniques. For example, in analyses of a single mediator without exposure-induced confounding, natural direct and indirect effects can be estimated using a regression-based approach with linear models for both the outcome and mediator, a simulation approach that accommodates nonlinear models, and a weighting estimator that requires only models for the exposure. While these parametric methods are more practical than nonparametric estimation, they are not without limitations. In particular, the consistency of these estimators depends not only on the identification assumptions for their corresponding estimands but also on the correct specification of all requisite models. If any of these models are misspecified, the resulting estimates can be biased.

This chapter introduces a class of robust estimation methods for analyzing causal mediation that protect against bias due to model misspecification. To understand the properties of these methods, it may be helpful to briefly review several key concepts in estimation theory. In Chapter 3, we explained that an estimator is consistent if it converges in probability to the target estimand as the sample size increases. Given appropriate identifying assumptions and correct model misspecification, the parametric estimators discussed in previous chapters are all consistent. In fact, they are $\sqrt{n}$*-consistent* and *asymptotically normal.* This means that when scaled by the square root of the sample size, the deviation of the estimator from its target estimand converges to a normal distribution with mean zero and a finite variance as the sample size increases. Formally, an estimator $\hat{\theta}$ is $\sqrt{n}$-consistent for a target estimand θ if $\sqrt{n}\left(\hat{\theta}-\theta\right)$ is bounded in probability. In other words, for every $\epsilon > 0$, there is a $\nu > 0$ such that $P\left(|\sqrt{n}(\hat{\theta}-\theta)| > \nu\right) < \epsilon$ for all n. If a $\sqrt{n}$-consistent estimator $\hat{\theta}$ is also asymptotically normal, then $\sqrt{n}(\hat{\theta}-\theta) \overset{d}{\to} N(0,\sigma^2)$, where $\overset{d}{\to}$ denotes convergence in distribution and σ^2 is called the *asymptotic variance* of $\hat{\theta}$.

Compared with the parametric estimators discussed previously, the approach we introduce in this chapter has two distinctive features. First, for each estimand, it leverages nonparametric efficiency theory to construct an optimal estimator (Hines et al. 2022; Tsiatis 2006). This estimator involves several "nuisance" functions that need to be estimated from the sample data. These are functions that are not of primary scientific interest but are necessary for estimating the target parameter that is. An estimator derived from nonparametric efficiency theory is optimal because, if its nuisance functions are correctly specified and consistently estimated, the estimator will not only be $\sqrt{n}$-consistent and asymptotically normal but also *nonparametrically efficient*. This means that the estimator has the smallest asymptotic variance that can possibly be achieved with a nonparametric model. Moreover, the estimator is often *robust*, meaning that it will remain $\sqrt{n}$-consistent and asymptotically normal even when some of its nuisance functions are misspecified.

For example, as we outline below, an estimator for the average total effect derived from nonparametric efficiency theory involves two nuisance functions: a model for the conditional mean of the outcome given the exposure and baseline confounders, and a model for the conditional probability of the exposure given the baseline confounders. The estimator is "doubly robust" because it will be consistent if either the outcome model or the exposure model is correctly specified, but not necessarily both. Throughout the chapter, we introduce similar estimators for natural, interventional, and path-specific effects.

Second, the other distinctive feature of the approach we introduce in this chapter is its compatibility with machine learning methods (Chernozhukov et al. 2018; van der Laan and Rose 2011). Machine learning involves using an adaptive fitting procedure to model different types of functions, employing algorithms that learn patterns and relationships directly from the data. This inductive, algorithmic process of learning from data is quite powerful because it can accurately approximate many different functions without requiring the analyst to specify their form (Hastie et al. 2009).

Recognizing the vulnerability of parametric estimators to bias from model misspecification, social scientists are increasingly turning to machine learning methods to improve the robustness of their analyses (Athey and Imbens 2019; Brand et al. 2023; Grimmer et al. 2021). However, naive application of machine learning methods can result in suboptimal estimators that suffer from nontrivial bias in finite samples and converge to their target estimands at a rate slower than the square root of the sample size. These drawbacks can be mitigated by integrating machine learning with a carefully constructed estimator. In particular, for many estimands, if machine learning methods are used to estimate the nuisance functions of an estimator derived from nonparametric efficiency theory, it will be $\sqrt{n}$-consistent, asymptotically normal, and nonparametrically efficient under fairly general conditions. Because machine learning methods typically do not impose strong parametric constraints on the form of these nuisance functions, this approach adds another layer of robustness without sacrificing efficiency.

The rest of the chapter is organized as follows. First, we review the problem of model misspecification in analyses of average total effects, introduce a doubly robust estimator for the total effect, and explain why this estimator is well-suited for the application of machine learning methods. Next, we expand on this foundation by introducing a series of similar estimators for natural direct and indirect effects, interventional direct and indirect effects, and path-specific effects. Each of these estimators is multiply robust and also amenable for use with machine learning. Throughout this discussion, we continue to use data from the NLSY to illustrate key concepts and methods. We also present an additional empirical example based on data from a recent study on the multigenerational effects of political violence among Crimean Tatars (Lupu and Peisakhin 2017). Stata and R codes for implementing these analyses are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch6`, and the footnotes to all tables and figures include hyperlinks to the specific scripts and data files used to generate them.

6.1 Robust Estimation of Average Total Effects

In this section, we review the problem of model misspecification in analyses of average total effects, with a particular focus on parametric estimators that employ either regression imputation or inverse probability weighting. We then introduce a doubly robust estimator for the total effect that integrates both imputation and weighting. We discuss its unique properties and show how it can be implemented using machine learning methods.

6.1.1 Limitations of Parametric Estimators

The first step in almost all causal mediation analysis is to identify and estimate the average total effect. For an exposure D and an outcome Y, the total effect is defined as follows:

$$ATE\left(d, d^*\right) = \mathbb{E}\left[Y\left(d\right) - Y\left(d^*\right)\right], \tag{6.1}$$

where $Y\left(d\right)$ denotes the potential outcome under exposure to d and $Y\left(d^*\right)$ is defined analogously. This effect represents the expected difference in the outcome if individuals had experienced level d rather than d^* of the exposure. As outlined in Chapter 3, the total effect can be identified nonparametrically with the following function of observable data under assumptions (a.i) to (a.iii):

$$ATE\left(d, d^*\right) = \sum_{c}\left(\mathbb{E}\left[Y|c, d\right] - \mathbb{E}\left[Y|c, d^*\right]\right) P\left(c\right), \tag{6.2}$$

where C denotes a set of baseline confounders. In principle, the total effect can also be estimated nonparametrically by stratifying the data by all possible levels of the baseline confounders, evaluating the difference in stratum-specific means of the outcome between those with level d versus d^* of the exposure, and then computing a weighted average of these stratum-specific differences in means, with weights equal to the relative size of each stratum. However, this approach to estimation is not generally viable unless the exposure and baseline confounders are low-dimensional and discrete.

To circumvent this challenge, we could adopt one of several different parametric approaches for estimating the total effect. For simplicity, we now focus on estimating the mean of the potential outcomes under an arbitrary level of the exposure d, which can be denoted as $\psi_d = \mathbb{E}\left[Y\left(d\right)\right]$. If we have an estimator for ψ_d, denoted by $\hat{\psi}_d$, the total effect can be estimated by evaluating it twice–once for level d and again for level d^* of the exposure–and then by taking the difference between the resulting quantities, $\hat{\psi}_d$ and $\hat{\psi}_{d^*}$.

Under assumptions (a.i) to (a.iii) from Chapter 3, ψ_d can be identified using an iterated expectation as follows:

$$\psi_d = \sum_{c} \mathbb{E}\left[Y|c, d\right] P\left(c\right) = \mathbb{E}\left[\mathbb{E}\left[Y|c, d\right]\right]. \tag{6.3}$$

This expression indicates that we can estimate ψ_d using a regression-imputation approach, which is implemented through the following series of steps:

1. **Fit a model for the outcome.** That is, fit a model for the conditional mean of the outcome given the exposure and baseline confounders, denoted by $\mu_D\left(C\right) = \mathbb{E}\left[Y|C, D\right]$;

2. **Impute the potential outcomes.** Use the fitted model for the outcome to predict the mean of the

potential outcomes under exposure d by setting $D = d$ for all sample members and then computing predicted values, denoted by $\hat{\mu}_d(C) = \hat{\mathbb{E}}[Y|C, D = d]$.

3. **Compute an estimate.** To estimate ψ_d, calculate the sample average of the predicted outcomes, which can be expressed as follows:
$$\hat{\psi}_d^{ri} = \frac{1}{n}\sum \hat{\mu}_d(C), \tag{6.4}$$
where n is the sample size and the "ri" superscript indicates that this procedure yields a regression-imputation estimator.

By extension, a corresponding estimator for the total effect is given by $\widehat{ATE}(d, d^*)^{ri} = \hat{\psi}_d^{ri} - \hat{\psi}_{d^*}^{ri}$. If the outcome model in step 1 is linear and additive, then the regression imputation estimator will just equal the coefficient on the exposure in this model multiplied by $(d - d^*)$, but many other other specifications are possible.

Under assumptions (a.i) to (a.iii), Equation 6.3 can also be expressed as follows:

$$\psi_d = \mathbb{E}\left[\frac{I(D = d)Y}{P(d|C)}\right],$$

where $I(\cdot)$ is an indicator function equal to 1 when its argument is true, and 0 otherwise. This expression indicates that we can estimate ψ_d with inverse probability weighting as well, which is implemented through the following steps:

1. **Fit a model for the exposure.** That is, fit a model for the conditional probability of the exposure given the baseline confounders, denoted by $\pi_D(C) = P(D|C)$.

2. **Compute predicted probabilities of exposure.** For each sample member, use the fitted model for the exposure to predict the probability of exposure to d , denoted by $\hat{\pi}_d(C) = \hat{P}(D = d|C)$.

3. **Construct inverse probability weights.** Among sample members for whom $D = d$, compute a set of inverse probability weights given by $\hat{w} = {}^1/_{\hat{\pi}_d(C)}$.

4. **Compute an estimate.** To estimate ψ_d, calculate a weighted average of the outcome among sample members for whom $D = d$, with weights equal to $\hat{w}$. This can be achieved by evaluating the following expression:
$$\hat{\psi}_d^{ipw} = \frac{\sum I(D = d)\hat{w}Y}{\sum I(D = d)\hat{w}}, \tag{6.5}$$
where the "ipw" superscript denotes an estimator based on inverse probability weights.

As before, a corresponding estimator for the total effect is given by $\widehat{ATE}(d, d^*)^{ipw} = \hat{\psi}_d^{ipw} - \hat{\psi}_{d^*}^{ipw}$.

Comparing these two procedures illustrates how regression imputation and inverse probability weighting involve modeling different parts of the data distribution. Regression imputation requires a correctly specified model for the conditional mean of the outcome, while weighting depends on a correctly specified model for the conditional probability of exposure. If assumptions (a.i) to (a.iii) are met and the requisite model is correctly specified, these estimators are $\sqrt{n}$-consistent and asymptotically normal.

However, correctly specifying either the exposure or the outcome model is challenging in practice, especially when there are many baseline confounders. Misspecification of the outcome model leads to bias in the regression-imputation estimator; similarly, the weighting estimator is biased if the exposure model is misspecified. Both types of misspecification are common in the social sciences, where researchers rarely possess knowledge about the functional form through which these different variables are related, often leading to ritualistic adoption of arbitrary modeling conventions.

In an effort to counteract the risk of misspecification bias, researchers have recently turned to various machine learning methods to fit the models necessary for constructing imputation and weighting estimates (e.g., Hill 2011; Lee et al. 2010; McCaffrey et al. 2004). These methods include high-dimensional GLMs with the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996), classification and regression trees (CARTs; Maimon and Rokach 2014), gradient boosted CARTs (Friedman 2001), random forests (Breiman 2001), neural networks (Goodfellow et al. 2016), and stacking algorithms known as "super learners" (van der Laan et al. 2007). For a comprehensive introduction to these and other machine learning methods, see Hastie et al. (2009). The purported advantage of using these techniques over traditional methods (e.g., linear or logistic regression) lies in their ability to flexibly learn the form of the exposure or outcome model inductively from the data, thereby imposing much weaker constraints on the structure of these models and reducing the potential for misspecification.

Nevertheless, despite the capacity of these methods to accurately learn models for the exposure and outcome, they may not always outperform more traditional modeling approaches when used to implement regression imputation or weighting estimators. This is because machine learning algorithms are primarily designed to minimize prediction error, not to estimate causal effects. For example, the LASSO tends to fit models composed of the covariates that most effectively predict the outcome, but this subset of predictors may not be optimal for estimating the total effect of an exposure. If the LASSO excludes covariates that strongly predict the exposure, even if their predictive power with respect to the outcome is modest, significant bias can arise in regression imputation estimates of the total effect (Belloni et al. 2014). In general, machine learning methods tend to minimize prediction error by balancing the amount of bias and variance in their predictions. Although the bias in these predictions typically decreases as the sample size grows, this reduction often occurs slowly. As a result, naive use of machine learning methods with the imputation or weighting approach can lead to suboptimal estimators of causal effects that are biased in finite samples and converge to their target estimands at comparatively slow rates.

6.1.2 Doubly Robust Estimation

In this section, we introduce a third estimator for ψ_d, which is more resilient to model misspecification and more compatible with machine learning methods compared to regression imputation and inverse probability weighting. This estimator combines elements of both approaches by incorporating a model for the conditional mean of the outcome given the exposure and baseline confounders and a model for the conditional probability of the exposure given the baseline confounders. Despite requiring both an outcome model and an exposure model, this estimator is more robust to misspecification than either regression imputation or inverse probability weighting alone. It achieves this by only necessitating correct specification of *either* the outcome model *or* the exposure model, but not necessarily both (Scharfstein et al. 1999). This feature provides two opportunities to obtain consistent estimates of ψ_d: when the outcome model is correctly specified, or when the exposure model is correctly specified. Because it possesses two different pathways to consistency, this estimator is commonly described as "doubly robust" (DR).

Under assumptions (a.i) to (a.iii), Equation 6.3 can also be expressed as follows:

$$\psi_d = \mathbb{E}\left[\mathbb{E}\left[Y|C, D=d\right] + \frac{I\left(D=d\right)}{P\left(d|C\right)}\left(Y - \mathbb{E}\left[Y|C, D\right]\right)\right]. \tag{6.6}$$

This expression differs from Equation 6.3 in that it contains an additional term, $\left(I(D=d)/P(d|C)\right)\left(Y - \mathbb{E}\left[Y|C, D\right]\right)$, with mean zero, indicating that ψ_d can be estimated using a combination of regression imputation and weighting. Specifically, it indicates that we can estimate ψ_d as follows:

$$\hat{\psi}_d^{dr} = \frac{1}{n}\sum \hat{S}_d^{dr}, \tag{6.7}$$

$$\text{where} \quad \hat{S}_d^{dr} = \hat{\mu}_d\left(C\right) + \frac{I\left(D = d\right)}{\hat{\pi}_d\left(C\right)}\left(Y - \hat{\mu}_D\left(C\right)\right), \tag{6.8}$$

In this expression, $\hat{\mu}_d\left(C\right)$ are the imputed outcomes from step 2 of the regression imputation estimator, and $\hat{\pi}_d\left(C\right)$ are the predicted probabilities of exposure from step 2 of the weighting estimator. The superscript "dr" signifies that this estimator is doubly robust, meaning that it is consistent for ψ_d if either the model used to compute $\hat{\mu}_d\left(C\right)$ is correctly specified or if the model used to compute $\hat{\pi}_d\left(C\right)$ is correctly specified.

To appreciate why $\hat{\psi}_d^{dr}$ is doubly robust, first note that $Y - \hat{\mu}_D\left(C\right)$ is a residual term from the outcome model. If the outcome model is correctly specified, these residuals are uncorrelated in expectation with any function of the predictors in this model (i.e., the exposure and baseline confounders), including the function $I(D=d)/\hat{\pi}_d(C)$. Thus, the mean of $\left(I(D=d)/\hat{\pi}_d(C)\right)\left(Y - \hat{\mu}_D\left(C\right)\right)$ will converge to zero as the sample size increases, and as a result, $\hat{\psi}_d^{dr}$ will converge to $\frac{1}{n}\sum \hat{\mu}_d\left(C\right)$, which is equivalent to the regression imputation estimator for ψ_d. When the outcome model is correctly specified, regression imputation is consistent, and by extension, so is the DR estimator.

Second, note that $\hat{S}_d^{dr}$ can also be expressed as follows:

$$\hat{S}_d^{dr} = \frac{I\left(D = d\right)Y}{\hat{\pi}_d\left(C\right)} + \frac{\hat{\mu}_d\left(C\right)}{\hat{\pi}_d\left(C\right)}\left(I\left(D = d\right) - \hat{\pi}_d\left(C\right)\right),$$

where $I\left(D = d\right) - \hat{\pi}_d\left(C\right)$ is a residual term from the model for the conditional probability of the exposure. If the exposure model is correctly specified, these residuals are uncorrelated in expectation with any function of the predictors in this model (i.e., the baseline confounders), including the function $\hat{\mu}_d(C)/\hat{\pi}_d(C)$. Thus, the mean of $\left(\hat{\mu}_d(C)/\hat{\pi}_d(C)\right)\left(I\left(D = d\right) - \hat{\pi}_d\left(C\right)\right)$ will also converge to zero as the sample size increases, and as a result, $\hat{\psi}_d^{dr}$ will converge to $\frac{1}{n}\sum I(D=d)Y/\hat{\pi}_d(C)$, which is equivalent to an inverse probability weighting estimator for ψ_d. When the exposure model is correctly specified, the weighting estimator is consistent, and by extension, so is the DR estimator.

In sum, $\hat{\psi}_d^{dr}$ is doubly robust–consistent if either the outcome model or the exposure model is correctly specified. Moreover, this estimator will also be $\sqrt{n}$-consistent and asymptotically normal provided that $\mu_D\left(C\right)$ and $\pi_D\left(C\right)$ are estimated using parametric models and at least one of them is correctly specified. If both the outcome and exposure models are correctly specified, the performance of the DR estimator will be even better. In this case, it will also be nonparametrically efficient.

6.1.3 Debiased Machine Learning

Compared with regression imputation and inverse probability weighting, another advantage of the DR estimator is its compatibility with the use of machine learning methods for estimating $\mu_D\left(C\right)$ and $\pi_D\left(C\right)$ (Chernozhukov et al. 2018; van der Laan and Rose 2011). This compatibility stems from the fact that the bias in $\hat{\psi}_d^{dr}$ is governed by the product of the bias in $\hat{\mu}_D\left(C\right)$ and the bias in $\hat{\pi}_D\left(C\right)$. This multiplicative form implies that when one of these two bias terms converges to zero, the bias in $\hat{\psi}_d^{dr}$ converges to zero as well, hence its double robustness. It also implies that the DR estimator can achieve $\sqrt{n}$-consistency even if the models used to estimate $\mu_D(C)$ and $\pi_D(C)$ are not $\sqrt{n}$-consistent themselves, as is often the case with data-adaptive machine learning methods. Owing to the multiplicative form of its bias, the DR estimator will be $\sqrt{n}$-consistent for ψ_d as long as the product of the convergence rates for the outcome model and the exposure model is faster than $\sqrt{n}$ (e.g., if both converge at a rate faster than $n^{1/4}$). This condition can be achieved by many machine learning methods, including the LASSO (Bickel et al. 2009), CARTs and random forests (Wager and Walther 2015), and neural networks (Chen and White 1999; Farrell et al. 2021).

When using machine learning methods to fit the outcome and exposure models, it is also advisable to implement them in conjunction with a sample splitting procedure, where a portion of the data is used to fit the estimator's nuisance functions (i.e., the exposure and outcome models) and another portion is used to compute the DR estimator. Such procedures remove any "overfitting bias" that may afflict machine learning methods when used to estimate causal effects. This type of bias can arise if machine learning models, especially complex ones, fit the estimator's nuisance functions too closely. When the same data are used to estimate both the nuisance functions and the causal parameter of interest, this "overfitting" can introduce bias into the final estimator. By ensuring that the data used to estimate the nuisance functions is separate from the data used to estimate the target parameter, sample splitting mitigates the bias that can arise when the same data are used for both purposes.

To this end, we recommend using a *repeated cross-fitting* procedure (Chernozhukov et al. 2018), which can be implemented as follows:

1. **Partition the sample data.** Randomly divide the sample data into $4 \leq J \leq 10$ partitions, or subsamples, denoted by $S_1, S_2, \ldots, S_J$.

2. **Compute estimates for each partition.** That is, for $k = 1, 2, \ldots, J$,

   (a) Fit machine learning models for the outcome and exposure using all the sample partitions except for S_k.

   (b) Then, with the models fit to data from all but the k^{th} sample partition, compute a DR estimate for ψ_d now using only the data from partition S_k. Let $\hat{\psi}_d^{dr,k}$ denote this partition-specific estimate.

3. **Combine estimates across partitions.** Compute a final estimate for ψ_d by averaging the partition-specific estimates, $\hat{\psi}_d^{dr,k}$ for $k = 1, 2, \ldots, J$, together.

Using the DR estimator in conjunction with repeated cross-fitting and machine learning models for the outcome and exposure is an instance of what Chernozhukov et al. (2018) call "debiased machine learning" (DML). An estimator for ψ_d based on DML will be $\sqrt{n}$-consistent, asymptotically normal, and nonparametrically efficient if assumptions (a.i) to (a.iii) are satisfied and provided that the product of the convergence rates for the outcome and exposure models is faster than the square root of the sample size.

6.1.4 Statistical Inference

We have thus far focused on point estimation of ψ_d and the average total effect. To compute inferential statistics based on the regression imputation or weighting approach, the nonparametric bootstrap can be used, following the same procedures outlined in Chapter 3. Alternatively, when DML is used and the product of the convergence rates for the exposure and outcome models is faster than $\sqrt{n}$, the resulting estimator is asymptotically normal with a sampling variance that can be estimated from the following expression:

$$\widehat{\mathrm{Var}}\left(\hat{\psi}_d^{dr}\right) = \frac{1}{n}\widehat{\mathrm{Var}}\left(\hat{S}_d^{dr}\right). \tag{6.9}$$

where $\mathrm{Var}(\,\cdot\,)$ represents the variance, "hats" distinguish estimates from estimands, and $\hat{S}_d^{dr}$ is defined as in Equation 6.8. Similarly, the sampling variance of a corresponding estimator for the total effect, $\widehat{ATE}\left(d, d^*\right)^{dr} = \hat{\psi}_d^{dr} - \hat{\psi}_{d^*}^{dr}$, can be estimated as follows:

$$\widehat{\mathrm{Var}}\left(\widehat{ATE}\left(d, d^*\right)^{dr}\right) = \frac{1}{n}\widehat{\mathrm{Var}}\left(\hat{S}_d^{dr} - \hat{S}_{d^*}^{dr}\right). \tag{6.10}$$

Taking the square root of these expressions yields an analytic standard error for $\hat{\psi}_d^{dr}$ and $\widehat{ATE}\left(d, d^*\right)^{dr}$, respectively, which can then be used to construct confidence intervals and compute p-values with the normal distribution, following standard inferential procedures. For example, to compute a 95% confidence interval for the total effect, we would compute $\widehat{ATE}\left(d, d^*\right)^{dr} \pm 1.96 \times \sqrt{\widehat{\mathrm{Var}}\left(\widehat{ATE}\left(d, d^*\right)^{dr}\right)}$.

When parametric models are used to estimate $\hat{\psi}_d^{dr}$ and $\widehat{ATE}\left(d, d^*\right)^{dr}$, rather than DML, these expressions for the sampling variance will remain valid only if the outcome and exposure models are both correctly specified. Otherwise, they will not in general provide accurate estimates of the sampling variance. Thus, we recommend computing inferential statistics with Equations 6.9 and 6.10 only when using DML. When implementing the DR estimator with parametric models, we recommend using the nonparametric bootstrap instead, to avoid drawing invalid inferences.

6.1.5 An Empirical Illustration: The Effect of College Attendance on Depression

To illustrate, we now apply regression imputation, weighting, DR estimation, and DML to analyze the total effect of college attendance on depression, using data from the NLSY. As discussed previously, regression imputation involves fitting a model for the conditional mean of the outcome given the exposure and baseline confounders, while inverse probability weighting involves fitting a model for the conditional probability of the exposure given baseline confounders. DR estimation involves fitting both of these models. For the regression imputation, weighting, and parametric DR estimators, we use a linear model for the outcome (CES-D scores) and a logit model for the exposure (college attendance).[1] For the DML approach, we use repeated cross-fitting with $J = 5$ partitions and fit both the outcome and exposure models using a super learner composed of the LASSO and a random forest. For the weighting, DR, and DML estimators, we censor their weights at the 1st and 99th percentiles for improved precision.

Table 6.1 presents estimates of the total effect and the associated means of the potential outcomes (ψ_0 and ψ_1). The regression-imputation estimates suggest that attending college reduces CES-D scores by .07 standard deviations, on average, while the estimate from the weighting approach is considerably more pronounced. This discrepancy raises concerns about potential misspecification of either the outcome model, the exposure model, or both. As a result, we might prioritize the DR estimates, especially those based on DML.

The parametric DR estimator indicates that college attendance lowers CES-D scores by about .144 stan-

[1] To be consistent with the DR estimator, we implement the weighting estimator with inverse probability weights that have not been normalized, although results based on normalized weights, as in Equation 6.5, are similar.

Table 6.1: Estimates for the Average Total Effect of College Attendance on Depression (CES-D scores) from the NLSY.

Estimand	Regression Imputation	Weighting	Parametric DR	DML
ψ_0	$.020\ [-.020, .063]$	$.039\ [-.002, .083]$	$.032\ [-.010, .078]$	$.037\ [-.009, .083]$
ψ_1	$-.049\ [-.113, .016]$	$-.125\ [-.215, -.025]$	$-.112\ [-.205, -.007]$	$-.102\ [-.201, -.004]$
$ATE\,(1, 0)$	$-.070\ [-.149, .012]$	$-.164\ [-.268, -.057]$	$-.144\ [-.250, -.031]$	$-.139\ [-.247, -.030]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals. For the RI, IPW, and DR estimates, the confidence intervals were computed using the nonparametric bootstrap with $B = 2000$ replications. For the DML estimates, confidence intervals are constructed using ±1.96 times the analytical standard errors described in Section 6.1.4. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch6/table_6-1`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

dard deviations, on average, positioning this estimate in between those obtained from regression imputation and weighting, albeit closer to the latter. The large discrepancy between the regression-imputation and DR estimates ($-.07$ versus $-.144$) in particular suggests that the outcome model may be misspecified, as its correct specification should lead these estimates to align more closely in large samples.

Results from the DML approach are quite similar to those based on parametric DR estimation. According to the DML estimate for the total effect, attending college reduces CES-D scores by about .139 standard deviations, and its 95% confidence interval does not span zero. Because this approach does not rely on any specific parametric model for either the outcome or the exposure, DML is less likely to be distorted by bias from model misspecification, compared with the other approaches. Links to the code and data used for this analysis are provided in the table footnote.

6.2 Robust Estimation of Natural Direct and Indirect Effects

In this section, we introduce robust methods for estimating natural direct and indirect effects in the presence of baseline confounding only. Specifically, we present two "multiply robust" (MR) estimators for these effects. The first estimator requires fitting models for the outcome, the mediator, and the exposure (Tchetgen Tchetgen and Shpitser 2012), while the second requires fitting two outcome models and two exposure models (Farbmacher et al. 2022; Zheng and van der Laan 2012; Zhou 2022). These estimators are multiply robust in that they remain consistent for the effects of interest under multiple different forms of model misspecification. Moreover, both estimators are compatible with DML, which offers additional protection against bias arising from misspecification.

As detailed in Chapter 3, the average total effect of an exposure D on an outcome Y can be separated into natural direct and indirect effects operating through a mediator M via the following decomposition:

$$ATE\,(d, d^*) = \underbrace{\mathbb{E}\left[Y\left(d, M\left(d^*\right)\right) - Y\left(d^*, M\left(d^*\right)\right)\right]}_{NDE(d,d^*)} + \underbrace{\mathbb{E}\left[Y\left(d, M\left(d\right)\right) - Y\left(d, M\left(d^*\right)\right)\right]}_{NIE(d,d^*)}.$$

Let $\psi_{d^*,d} = \mathbb{E}\left[Y\left(d, M\left(d^*\right)\right)\right]$ represent the marginal mean of the nested potential outcomes, such that the natural effects decomposition can be equivalently expressed as follows:

$$ATE\left(d, d^*\right) = \underbrace{\psi_{d^*,d} - \psi_{d^*,d^*}}_{NDE(d,d^*)} + \underbrace{\psi_{d,d} - \psi_{d^*,d}}_{NIE(d,d^*)}.$$

To identify and estimate this decomposition, it suffices to identify and estimate $\psi_{d^*,d}$ for any combination of d and d^*. Then, if we have an estimator for $\psi_{d^*,d}$, denoted by $\hat{\psi}_{d^*,d}$, the natural direct and indirect effects can be estimated as $\hat{\psi}_{d^*,d} - \hat{\psi}_{d^*,d^*}$ and $\hat{\psi}_{d,d} - \hat{\psi}_{d^*,d}$, respectively.

Under assumptions (c.i) to (c.vi) from Chapter 3, $\psi_{d^*,d}$ can be identified using an iterated expectation as follows:

$$\begin{aligned} \psi_{d^*,d} &= \sum_{c,m} \mathbb{E}\left[Y|c,d,m\right] P\left(m|c,d^*\right) P\left(c\right) \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y|C,d,M\right]|C,d^*\right]\right]. \end{aligned} \tag{6.11}$$

This identification formula can also be expressed as:

$$\begin{aligned} \psi_{d^*,d} = \mathbb{E}\Big[& \underbrace{\frac{I\left(D=d\right)}{P\left(d|C\right)} \left(\frac{P\left(M|C,d^*\right)}{P\left(M|C,d\right)}\right)\left(Y - \mathbb{E}\left[Y|C,d,M\right]\right)}_{\text{term 1}} \\ & + \underbrace{\frac{I\left(D=d^*\right)}{P\left(d^*|C\right)}\left(\mathbb{E}\left[Y|C,d,M\right] - \sum_m \mathbb{E}\left[Y|C,d,m\right] P\left(m|C,d^*\right)\right)}_{\text{term 2}} \\ & + \underbrace{\sum_m \mathbb{E}\left[Y|C,d,m\right] P\left(m|C,d^*\right)}_{\text{term 3}} \Big], \end{aligned} \tag{6.12}$$

where the last term in this expression is equal to $\mathbb{E}\left[\mathbb{E}\left[Y|C,d,M\right]|C,d^*\right]$. Equation 6.12 therefore differs from Equation 6.11 in that it contains two additional terms–term 1 and term 2–each of which has a mean of zero.

This version of the identification formula suggests an estimator for $\psi_{d^*,d}$ with the following form:

$$\hat{\psi}^{mr_1}_{d^*,d} = \frac{1}{n}\sum \hat{S}^{mr_1}_{d^*,d}, \tag{6.13}$$

$$\begin{aligned} \text{where} \quad \hat{S}^{mr_1}_{d^*,d} =& \frac{I\left(D=d\right)}{\hat{\pi}_d\left(C\right)}\left(\frac{\hat{P}\left(M|C,d^*\right)}{\hat{P}\left(M|C,d\right)}\right)\left(Y - \hat{\mu}_d\left(C,M\right)\right) \\ & + \frac{I\left(D=d^*\right)}{\hat{\pi}_{d^*}\left(C\right)}\left(\hat{\mu}_d\left(C,M\right) - \sum_m \hat{\mu}_d\left(C,m\right)\hat{P}\left(m|C,d^*\right)\right) \\ & + \sum_m \hat{\mu}_d\left(C,m\right)\hat{P}\left(m|C,d^*\right). \end{aligned} \tag{6.14}$$

In this expression, $\pi_D\left(C\right)$ denotes the conditional probability of the exposure D given the baseline confounders C, $P\left(M|C,D\right)$ denotes the conditional probability of the mediator M given the exposure D and the baseline confounders C, and $\mu_D\left(C,M\right)$ denotes the conditional mean of the outcome under exposure D given the mediator M and baseline confounders C. The “hats” on these quantities signal that they are estimates. For example, $\hat{\pi}_d\left(C\right)$ denotes the predicted probability that an individual is exposed to d given their baseline confounders C, $\hat{P}\left(M|d,C\right)$ denotes the predicted probability of an individual's observed mediator under exposure d and given their baseline confounders C, and $\hat{\mu}_d\left(C,M\right)$ denotes the predicted value for the outcome under exposure d given an individual's baseline confounders C and their observed value on the mediator M.

The estimator $\hat{\psi}^{mr_1}_{d^*,d}$ requires three different models, which constitute its nuisance functions: one for the conditional probability of the exposure $\pi_D\left(C\right)$; another for the conditional probability of the mediator $P\left(M|C,D\right)$; and a third for the conditional mean of the outcome $\mu_D\left(C,M\right)$. This estimator is multiply robust, meaning that it remains consistent for $\psi_{d^*,d}$ even when one of these three models is misspecified. In other words, it is consistent when the exposure model and the mediator model are correctly specified, when the mediator model and the outcome model are correctly specified, or when the exposure model and the outcome model are correctly specified. If any one of these conditions hold, along with assumptions (c.i) to (c.vi), then $\hat{\psi}^{mr_1}_{d^*,d}$ is consistent for $\psi_{d^*,d}$. Because it possesses three different pathways to consistency, this estimator is sometimes described as "triply robust."

To appreciate why $\hat{\psi}^{mr_1}_{d^*,d}$ is triply robust, consider each of the following scenarios in turn. First, if the exposure and mediator models are correctly specified, all terms involving the outcome model in Equation 6.14 have an expected value of zero, such that $\hat{\psi}^{mr_1}_{d^*,d}$ converges to $\frac{1}{n}\sum{\left(I(D=d)\hat{P}(M|d^*,C)Y\right)}/{\hat{\pi}_d(C)\hat{P}(M|d,C)}$ as the sample size increases. This expression is similar to an estimator based on ratio-of-mediator-probability weighting, which we introduced in Appendix F, and it is consistent for $\psi_{d^*,d}$ because the requisite models for $\pi_D\left(C\right)$ and $P\left(M|C,D\right)$ are both correctly specified in this scenario.

Second, if the mediator and outcome models are correctly specified, the first two terms in Equation 6.14 have an expected value of zero, such that the MR estimator converges to $\frac{1}{n}\sum\left(\sum_m \hat{\mu}_d\left(C,m\right)\hat{P}\left(m|C,d^*\right)\right)$ as the sample size increases. This expression is similar to a regression-imputation estimator for the mean of the nested potential outcomes, which is consistent in this scenario because its requisite models for $P\left(M|C,D\right)$ and $\mu_D\left(C,M\right)$ are both correctly specified.

Finally, if the exposure and outcome models are correctly specified, all terms involving the mediator model in Equation 6.14 have an expected value of zero, such that $\hat{\psi}^{mr_1}_{d^*,d}$ converges to $\frac{1}{n}\sum{I(D=d^*)\hat{\mu}_d(C,M)}/{\hat{\pi}_{d^*}(C)}$ as the sample size increases. This expression is similar to an imputation-based weighting estimator for the mean of the nested potential outcomes. It is consistent in this scenario because the models for $\pi_D\left(C\right)$ and $\mu_D\left(C,M\right)$ are both correctly specified.

To summarize, $\hat{\psi}^{mr_1}_{d^*,d}$ is triply robust–that is, consistent when at least two of its three requisite models are correctly specified. Moreover, this estimator will also be $\sqrt{n}$-consistent and asymptotically normal provided that $\pi_D\left(C\right)$, $P\left(M|C,D\right)$, and $\mu_D\left(C\right)$ are estimated using parametric models and at least two of them are correctly specified. If all three models are correct, the MR estimator will be nonparametrically efficient as well.

The estimator $\hat{\psi}^{mr_1}_{d^*,d}$ requires fitting a model for the conditional probability of the mediator given the exposure and baseline confounders, denoted by $P\left(M|C,D\right)$. This is relatively easy to accomplish when the mediator is univariate and discrete with only a few values. For example, researchers could fit a conventional logit model when the mediator is binary, a multinomial logit model when the mediator is polytomous, or an ordered logit model when the mediator is ordinal. However, $\hat{\psi}^{mr_1}_{d^*,d}$ is difficult to use in applications where the mediator is multivariate, continuous, or discrete with more than a handful values. In these scenarios, estimates for the conditional probability (or density) of the mediator can be unstable and highly sensitive to misspecification.

For applications with a multivariate, continuous, or many-valued mediator, we can recast the ratio of mediator probabilities in term 1 of Equation 6.12 as odds ratios involving only a set of exposure probabilities. Specifically, using Bayes' rule, we can express ${}^{P(M|d^*,C)}/_{P(M|d,C)}$ as follows:

$$\frac{P\left(M|d^*,C\right)}{P\left(M|d,C\right)}=\frac{P\left(d^*|C,M\right)P\left(d|C\right)}{P\left(d|C,M\right)P\left(d^*|C\right)}.$$

In addition, we can also recast the quantity $\sum_m \mathbb{E}\left[Y|C,d,m\right]P\left(m|C,d^*\right)$ in terms 2 and 3 of Equation 6.12 as an iterated expectation, such that $\psi_{d^*,d}$ can be alternatively expressed as follows:

$$\begin{aligned}\psi_{d^*,d}=\mathbb{E}\Big[&\underbrace{\frac{I\left(D=d\right)}{P\left(d^*|C\right)}\left(\frac{P\left(d^*|C,M\right)}{P\left(d|C,M\right)}\right)\left(Y-\mathbb{E}\left[Y|C,d,M\right]\right)}_{\text{term 1}}\\&+\underbrace{\frac{I\left(D=d^*\right)}{P\left(d^*|C\right)}\left(\mathbb{E}\left[Y|C,d,M\right]-\mathbb{E}\left[\mathbb{E}\left[Y|C,d,M\right]|C,d^*\right]\right)}_{\text{term 2}}\\&+\underbrace{\mathbb{E}\left[\mathbb{E}\left[Y|C,d,M\right]|C,d^*\right]}_{\text{term 3}}\Big].\end{aligned}\tag{6.15}$$

Equation 6.15 differs from Equation 6.12 in that it no longer contains any terms involving the conditional probability of the mediator.

This recast version of the identification formula for $\psi_{d^*,d}$ suggests an alternative estimator with the following form:

$$\hat{\psi}^{mr_2}_{d^*,d}=\frac{1}{n}\sum\hat{S}^{mr_2}_{d,d^*},\tag{6.16}$$

$$\begin{aligned}\text{where}\quad\hat{S}^{mr_2}_{d^*,d}=&\frac{I\left(D=d\right)}{\hat{\pi}_{d^*}\left(C\right)}\left(\frac{\hat{\pi}_{d^*}\left(C,M\right)}{\hat{\pi}_{d}\left(C,M\right)}\right)\left(Y-\hat{\mu}_d\left(C,M\right)\right)\\&+\frac{I\left(D=d^*\right)}{\hat{\pi}_{d^*}\left(C\right)}\left(\hat{\mu}_d\left(C,M\right)-\hat{\nu}_{d^*}\left(C\right)\right)\\&+\hat{\nu}_{d^*}\left(C\right),\end{aligned}\tag{6.17}$$

where $\pi_D\left(C\right)$ and $\mu_D\left(C,M\right)$ are defined as before, $\pi_D\left(C,M\right)$ represents the conditional probability of the exposure D given the baseline confounders C and mediator M, and $\nu_D\left(C\right)$ represents the conditional mean of $\mu_d\left(C,M\right)$ under exposure D given the baseline confounders C. The estimator $\hat{\psi}^{mr_2}_{d^*,d}$ requires four different models. Specifically, it requires two exposure models–one for $\pi_D\left(C\right)$ and the other for $\pi_D\left(C,M\right)$. And it also requires two outcome models–one for $\mu_D\left(C,M\right)$ and another for $\nu_D\left(C\right)$. These are the estimator's nuisance functions. Although it requires fitting an extra nuisance function, $\hat{\psi}^{mr_2}_{d^*,d}$ is much easier to implement than $\hat{\psi}^{mr_1}_{d^*,d}$ in applications where the mediator is multivariate, continuous, or discrete with many values because it does not require modeling the mediator.

Moreover, this estimator is also multiply robust in that it remains consistent in the presence of several different forms of model misspecification. In particular, $\hat{\psi}^{mr_2}_{d^*,d}$ is consistent for $\psi_{d^*,d}$ under assumptions (c.i) to (c.vi), when at least one of the following conditions holds: the models for $\pi_D\left(C\right)$ and $\pi_D\left(C,M\right)$ are correctly specified, the models for $\pi_D\left(C\right)$ and $\mu_D\left(C,M\right)$ are correctly specified, or the models for $\mu_D\left(C,M\right)$ and $\nu_D\left(C\right)$ are correctly specified. This estimator is also sometimes described as "triply robust," as it too possesses three distinct pathways for acheiving consistency. In addition, $\hat{\psi}^{mr_2}_{d^*,d}$ is $\sqrt{n}$-consistent and asymptotically normal if the four models it requires are all specified parametrically and at least one of the three conditions outlined previously is met. If all four models are correctly specified, $\hat{\psi}^{mr_2}_{d^*,d}$ is also nonparametrically efficient.

Both of the MR estimators introduced above are compatible with DML. This is because the bias for each of these estimators is governed by a sum of bias products. The multiplicative form of the bias for $\hat{\psi}^{mr_1}_{d^*,d}$ and $\hat{\psi}^{mr_2}_{d^*,d}$ implies that they will remain $\sqrt{n}$-consistent as long as estimates for each of the models they require converge at a rate faster than $n^{1/4}$. As noted previously, faster-than-$n^{1/4}$ convergence can be achieved by many machine learning methods. When MR estimation is implemented via DML with algorithms that converge at sufficiently fast rates, both $\hat{\psi}^{mr_1}_{d^*,d}$ and $\hat{\psi}^{mr_2}_{d^*,d}$ will be $\sqrt{n}$-consistent, asymptotically normal, and nonparametrically efficient.

When using DML, the sampling variance of $\hat{\psi}^{mr_1}_{d^*,d}$ can be estimated with the following expression:

$$\widehat{\text{Var}}\left(\hat{\psi}^{mr_1}_{d^*,d}\right) = \frac{1}{n}\widehat{\text{Var}}\left(\hat{S}^{mr_1}_{d^*,d}\right). \tag{6.18}$$

By extension, the sampling variances of corresponding estimators for total, natural direct, and natural indirect effects can be estimated as follows:

$$\begin{aligned}
\widehat{\text{Var}}\left(\widehat{NDE}\left(d,d^*\right)^{mr_1}\right) &= \frac{1}{n}\widehat{\text{Var}}\big(\hat{S}^{mr_1}_{d^*,d} - \hat{S}^{mr_1}_{d^*,d^*}\big) \\
\widehat{\text{Var}}(\widehat{NIE}\left(d,d^*\right)^{mr_1}) &= \frac{1}{n}\widehat{\text{Var}}\big(\hat{S}^{mr_1}_{d,d} - \hat{S}^{mr_1}_{d^*,d}\big) \\
\widehat{\text{Var}}(\widehat{ATE}\left(d,d^*\right)^{mr_1}) &= \frac{1}{n}\widehat{\text{Var}}\big(\hat{S}^{mr_1}_{d,d} - \hat{S}^{mr_1}_{d^*,d^*}\big).
\end{aligned} \tag{6.19}$$

Similar expressions can be used to estimate the sampling variance of $\hat{\psi}^{mr_2}_{d^*,d}$ and its corresponding estimators for natural effects. The square root of these expressions yields an analytic standard error, which can then be used together with the normal distribution to obtain confidence intervals and p-values, following standard inferential procedures.

However, when MR estimation is implemented with parametric models, rather than DML, the variance formulas in Equations 6.18 and 6.19 will be invalid if any of the requisite models for the exposure, mediator, or outcome are misspecified. With this approach to MR estimation, we therefore recommend computing inferential statistics using the bootstrap procedures described in Chapter 3.

Using data from the NLSY, we applied these methods to estimate the total, natural direct, and natural indirect effects of college attendance on depression, as mediated by unemployment. Since unemployment status is binary, we implemented both $\hat{\psi}^{mr_1}_{d^*,d}$ and $\hat{\psi}^{mr_2}_{d^*,d}$. For each of these approaches, we computed parametric and DML estimates. The parametric estimates for $\hat{\psi}^{mr_1}_{d^*,d}$ were obtained from a logit model for $\pi_D\left(C\right)$, another logit model for $P\left(M|C,D\right)$, and a linear model for $\mu_D\left(C,M\right)$. The parametric estimates for $\hat{\psi}^{mr_2}_{d^*,d}$, on the other hand, were obtained from a logit model for $\pi_D\left(C\right)$, another logit model for $\pi_D\left(C,M\right)$, a linear model for $\mu_D\left(C,M\right)$, and another linear model for $\nu_D\left(C\right)$. Specifically, the last model for $\nu_D\left(C\right)$ was fit using a linear regression of the predicted outcomes $\hat{\mu}_d\left(C,M\right)$ on the exposure D and baseline confounders C. To obtain the DML estimates, we used repeated cross-fitting with $J = 5$ partitions and fit all requisite models using a super learner composed of the LASSO and random forests. In all these analyses, we included the same set of baseline confounders in C as described in previous chapters.

Table 6.2 presents results from this robust approach to estimation, with links to the code and data provided in the table footnote. The two different MR estimators yield highly consistent results. Both suggest

Table 6.2: Multiply Robust Estimates of the Total, Natural Direct, and Natural Indirect Effects of College Attendance on CES-D scores, as Mediated by Unemployment, from the NLSY.

Estimand	$\hat{\psi}^{mr_1}_{d^*,d}$		$\hat{\psi}^{mr_2}_{d^*,d}$	
	Parametric	DML	Parametric	DML
$ATE\,(1,0)$	$-.144$ $[-.253, -.036]$	$-.112$ $[-.208, -.016]$	$-.144$ $[-.253, -.035]$	$-.121$ $[-.216, -.025]$
$NDE\,(1,0)$	$-.136$ $[-.245, -.027]$	$-.109$ $[-.206, -.012]$	$-.136$ $[-.245, -.027]$	$-.116$ $[-.215, -.016]$
$NIE\,(1,0)$	$-.009$ $[-.025, .008]$	$-.003$ $[-.013, .007]$	$-.008$ $[-.026, .010]$	$-.005$ $[-.019, .009]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals. For the parametric estimates, confidence intervals were computed using the nonparametric bootstrap with $B = 2000$ replications. For the DML estimates, confidence intervals are constructed using ± 1.96 times the analytical standard errors described in Section 6.2. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch6/table_6-2`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

a sizable total effect of college attendance on depression, but a minimal role for unemployment in mediating this effect. According to the parametric estimates, for example, the total effect of college attendance on CES-D scores is about .144 standard deviations, but only about 6% of this effect ($.009/.144 \approx .06$) appears to operate through unemployment. The results remain essentially unchanged when MR estimation is implemented via DML, although estimates for the total and natural direct effects are somewhat smaller with this approach. Taken together, these findings suggest that our inferences about the mediating role of unemployment are highly robust to potential model misspecification.

6.3 Robust Estimation of Interventional Direct and Indirect Effects

In the presence of exposure-induced confounding, natural direct and indirect effects are not nonparametrically identified. As discussed in Chapter 4, an alternative set of estimands that can be used to assess mediation in this setting are the interventional direct effect and interventional indirect effect. In this section, we introduce a robust approach to estimating these effects (Benkeser and Ran 2021; Diaz et al. 2021). Similar to the MR estimators for natural effects, the MR estimators for interventional effects are consistent under several different forms of model misspecification, and they are also compatible with DML.

Recall that interventional direct and indirect effects of an exposure D on an outcome Y are formally defined as follows:

$$IDE\,(d, d^*) = \mathbb{E}\left[Y\left(d, \mathcal{M}\left(d^*|C\right)\right) - Y\left(d^*, \mathcal{M}\left(d^*|C\right)\right)\right]$$
$$IIE\,(d, d^*) = \mathbb{E}\left[Y\left(d, \mathcal{M}\left(d|C\right)\right) - Y\left(d, \mathcal{M}\left(d^*|C\right)\right)\right],$$

where $\mathcal{M}\,(d|C)$ denotes a value of the mediator randomly drawn from its distribution under exposure d given the baseline confounders C. The sum of the interventional direct and indirect effects is equal to the overall effect, which is given by the following expression:

$$OE\,(d, d^*) = \mathbb{E}\left[Y\left(d, \mathcal{M}\left(d|C\right)\right) - Y\left(d^*, \mathcal{M}\left(d^*|C\right)\right)\right].$$

To simplify notation, let $\phi_{d^*,d} = \mathbb{E}\left[Y\left(d, \mathcal{M}\left(d^*|C\right)\right)\right]$ represent the marginal mean of the randomized potential outcomes, such that the decomposition of the overall effect into direct and indirect components can be expressed as follows:

$$OE\left(d, d^*\right) = \underbrace{\phi_{d^*,d} - \phi_{d^*,d^*}}_{IDE(d,d^*)} + \underbrace{\phi_{d,d} - \phi_{d^*,d}}_{IIE(d,d^*)}.$$

To identify and estimate this decomposition, it suffices to identify and estimate $\phi_{d^*,d}$ for any combination of d and d^*. Then, if we have an estimator for $\phi_{d^*,d}$, denoted by $\hat{\phi}_{d^*,d}$, the interventional direct and indirect effects can be estimated as $\hat{\phi}_{d^*,d} - \hat{\phi}_{d^*,d^*}$ and $\hat{\phi}_{d,d} - \hat{\phi}_{d^*,d}$, respectively.

Under assumptions (e.i) to (e.v) from Chapter 4, $\phi_{d^*,d}$ can be identified with the following function of observable data:

$$\begin{aligned}\psi^{int}_{d^*,d} &= \sum_{c,m,l} \mathbb{E}\left[Y|c,d,l,m\right] P\left(l|c,d\right) P\left(m|c,d^*\right) P\left(c\right) \\ &= \mathbb{E}\left[\mathbb{E}\left[\sum_{l} \mathbb{E}\left[Y|C,d,l,M\right] P\left(l|C,d\right) |C, d^*\right]\right], \qquad (6.20)\end{aligned}$$

where L denotes an exposure-induced confounder of the mediator-outcome relationship. This identification formula can also be expressed as follows:

$$\begin{aligned}\phi_{d^*,d} = \mathbb{E}\Big[&\underbrace{\frac{I\left(D=d\right)}{P\left(d|C\right)} w\left(C,d,L,M\right)\left(Y - \mathbb{E}\left[Y|C,d,L,M\right]\right)}_{\text{term 1}} \\ &+ \underbrace{\frac{I\left(D=d\right)}{P\left(d|C\right)}\left(u\left(C,d,L\right) - \sum_{l} u\left(C,d,l\right) P\left(l|C,d\right)\right)}_{\text{term 2}} \\ &+ \underbrace{\frac{I(D=d^*)}{P\left(d^*|C\right)}\left(\sum_{l} \mathbb{E}\left[Y|C,d,l,M\right] P\left(l|C,d\right) - v\left(C,d^*\right)\right)}_{\text{term 3}} \\ &+ \underbrace{v\left(C,d^*\right)}_{\text{term 4}}\Big], \qquad (6.21)\end{aligned}$$

where $w\left(C,D,L,M\right)$, $u\left(C,D,L\right)$, and $v\left(C,D\right)$ are defined as

$$\begin{aligned}w\left(C,D,L,M\right) &= \frac{P\left(D|C\right) P\left(L|C,D\right) P\left(d^*|C,M\right)}{P\left(d^*|C\right) P\left(L|C,D,M\right) P\left(D|C,M\right)} \\ u\left(C,D,L\right) &= \mathbb{E}\left[\mathbb{E}\left[Y|C,D,L,M\right] w\left(C,D,L,M\right)|C,D,L\right] \\ v\left(C,D\right) &= \mathbb{E}\left[\sum_{l} \mathbb{E}\left[Y|C,d,l,M\right] P\left(l|C,d\right)|C,D\right].\end{aligned}$$

In this version of the identification formula, the fourth term, $v\left(C,d^*\right)$, is equal to $\mathbb{E}\left[\sum_l \mathbb{E}\left[Y|C,d,l,M\right] P\left(l|C,d\right)|C,d^*\right]$. Equation 6.21 therefore differs from Equation 6.20 in that it contains three additional terms–term 1, term 2, and term 3–that each have a mean of zero.

The identification formula in Equation 6.21 suggests the following estimator for $\phi_{d^*,d}$:

$$\begin{aligned}
\hat{\phi}^{mr}_{d^*,d} =& \frac{1}{n}\sum \hat{R}^{mr}_{d^*,d}, && (6.22)\\
\text{where}\quad \hat{R}^{mr}_{d^*,d} =& \frac{I\left(D=d\right)}{\hat{\pi}_d\left(C\right)}\hat{w}\left(C,d,L,M\right)\left(Y-\hat{\mu}_d\left(C,L,M\right)\right)\\
&+\frac{I\left(D=d\right)}{\hat{\pi}_d\left(C\right)}\left(\hat{u}\left(C,d,L\right)-\sum_l \hat{u}\left(C,d,l\right)\hat{P}\left(l|C,d\right)\right)\\
&+\frac{I\left(D=d^*\right)}{\hat{\pi}_{d^*}\left(C\right)}\left(\sum_l \hat{\mu}_d\left(C,l,M\right)\hat{P}\left(l|C,d\right)-\hat{v}\left(C,d^*\right)\right)\\
&+\hat{v}\left(C,d^*\right). && (6.23)
\end{aligned}$$

In this expression, $\pi_D\left(C\right)$ denotes the conditional probability of the exposure D given the baseline confounders C, $P\left(L|C,D\right)$ denotes the conditional probability of the exposure-induced confounder L given the exposure D and baseline confounders C, and $\mu_D\left(C,L,M\right)$ denotes the conditional mean of the outcome Y under exposure D given the mediator M, baseline confounders C, and exposure-induced confounder L. The "hats" over these quantities denote that they are estimates from sample data. The estimator $\hat{\phi}^{mr}_{d^*,d}$ thus requires fitting three models–one for $\pi_D\left(C\right)$, another for $P\left(L|C,D\right)$, and a third for $\mu_D\left(C,L,M\right)$–in order to estimate each of these quantities in turn.

In addition, $\hat{\phi}^{mr}_{d^*,d}$ also requires estimates of $w\left(C,D,L,M\right)$, $u\left(C,D,L\right)$, and $v\left(C,D\right)$. Computing these estimates necessitates fitting another four more models. Specifically, it requires a model for the conditional probability of the exposure given the mediator and baseline confounders, denoted by $\pi_D\left(C,M\right)$; a model for the conditional probability of the exposure-induced confounder given the exposure, mediator, and baseline confounders, denoted by $P\left(L|C,D,M\right)$; a model for the conditional mean of $\mathbb{E}\left[Y|C,D,L,M\right]w\left(C,D,L,M\right)$ given the exposure, baseline confounders, and the exposure-induced confounder, denoted by $u\left(C,D,L\right)$; and finally, a model for the conditional mean of $\sum_l \mathbb{E}\left[Y|C,d,l,M\right]P\left(l|C,d\right)$ given the exposure and baseline confounders, denoted by $v\left(C,D\right)$. When fitting these last two models, the dependent variables are constructed from estimates of $\mu_D\left(C,L,M\right)$, $w\left(C,D,L,M\right)$, and $P\left(L|C,D\right)$.

In total, the estimator $\hat{\phi}^{mr}_{d^*,d}$ requires seven different models, which constitute its nuisance functions. These include two exposure models for $\pi_D\left(C\right)$ and $\pi_D\left(C,M\right)$, two confounder models for $P\left(L|C,D\right)$ and $P\left(L|C,D,M\right)$, and three outcome models for $\mu_D\left(C,L,M\right)$, $u\left(C,D,L\right)$, and $v\left(C,D\right)$. This estimator is multiply robust in that it remains consistent under multiple forms of model misspecification. Specifically, $\hat{\phi}^{mr}_{d^*,d}$ is consistent under assumptions (e.i) to (e.v), provided that at least one of the following three conditions is met: the models for $P\left(L|C,D\right)$, $\pi_D\left(C\right)$, and $\mu_D\left(C,L,M\right)$ are correctly specified; the models for $P\left(L|C,D\right)$, $\mu_D\left(C,L,M\right)$, and $v\left(C,D\right)$ are correctly specified; or the models for $P\left(L|C,D\right)$, $P\left(L|C,D,M\right)$, $\pi_D\left(C\right)$, and $\pi_D\left(C,M\right)$ are correctly specified. If at least one of these three conditions holds and the nuisance functions are all estimated using parametric models, then $\hat{\phi}^{mr}_{d^*,d}$ will also be $\sqrt{n}$-consistent and asymptotically normal. Moreover, if all the nuisance functions are correctly modeled, this estimator will be nonparametrically efficient as well.

Similar to the MR estimators for natural effects outlined in the previous section, the bias of $\hat{\phi}^{mr}_{d^*,d}$ is governed by a sum of bias products. Consequently, it is also compatible with DML. When implemented using DML, $\hat{\phi}^{mr}_{d^*,d}$ will be $\sqrt{n}$-consistent, asymptotically normal, and nonparametrically efficient as long as the estimates obtained for each nuisance function converge at a rate faster than $n^{1/4}$. In this case, the

Table 6.3: Multiply Robust Estimates for the Interventional Effects of College Attendance on CES-D Scores, as Mediated by Household Income, from the NLSY.

Estimand	Parametric MR	DML
$OE\,(1,0)$	$-.142$ $[-.259, -.025]$	$-.114$ $[-.215, -.013]$
$IDE\,(1,0)$	$-.102$ $[-.215, .011]$	$-.067$ $[-.165, .032]$
$IIE\,(1,0)$	$-.040$ $[-.062, -.018]$	$-.047$ $[-.069, -.026]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals. For the parametric estimates, the confidence intervals were computed using the nonparametric bootstrap with $B = 2000$ replications. For the DML estimates, confidence intervals are constructed using ± 1.96 times the analytical standard errors described in Section 6.3. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch6/table_6-3`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

sampling variance of $\hat{\phi}^{mr}_{d^*,d}$ can be estimated as follows:

$$\widehat{\text{Var}}\left(\hat{\phi}^{mr}_{d^*,d}\right) = \frac{1}{n}\widehat{\text{Var}}\left(\hat{R}^{mr}_{d^*,d}\right). \tag{6.24}$$

By extension, the sampling variance of corresponding estimators for the interventional effects of interest can be estimated with the follow expressions:

$$\begin{aligned}
\widehat{\text{Var}}\left(\widehat{IDE}\,(d,d^*)^{mr}\right) &= \frac{1}{n}\widehat{\text{Var}}\left(\hat{R}^{mr}_{d^*,d} - \hat{R}^{mr}_{d^*,d^*}\right) \\
\widehat{\text{Var}}\left(\widehat{IIE}\,(d,d^*)^{mr}\right) &= \frac{1}{n}\widehat{\text{Var}}\left(\hat{R}^{mr}_{d,d} - \hat{R}^{mr}_{d^*,d}\right) \\
\widehat{\text{Var}}\left(\widehat{OE}\,(d,d^*)^{mr}\right) &= \frac{1}{n}\widehat{\text{Var}}\left(\hat{R}^{mr}_{d,d} - \hat{R}^{mr}_{d^*,d^*}\right).
\end{aligned} \tag{6.25}$$

Taking the square root of these variance estimates provides a set of analytic standard errors, which can be used to construct confidence intervals and to compute p-values from the normal distribution. When $\hat{\phi}^{mr}_{d^*,d}$ is implemented with parametric models rather than DML, the nonparametric bootstrap should be used to compute inferential statistics instead, as the variance formulas provided above are invalid if any of these models are misspecified.

We applied these methods to data from the NLSY to reanalyze whether household income mediates the effect of college attendance on depression, while adjusting for unemployment as a potential exposure-induced confounder of the mediator-outcome relationship. As in Chapter 4, we transformed household income using its natural logarithm and included the same set of baseline confounders used in our prior analyses of these data. MR estimates for interventional effects were computed using both parametric models and DML. The parametric estimates were obtained from linear models for the outcome (CES-D scores) and logit models for the exposure (college attendance) and the exposure-induced confounder (unemployment). The DML estimates were obtained via cross-fitting–on $J = 5$ partitions–and super learners composed of the LASSO and a random forest for all requisite models.

Links to the code and data for this analysis are provided in the footnote to Table 6.3, which presents the parametric MR and DML estimates for the interventional effects of college attendance. The parametric approach generated somewhat larger estimates for the overall and direct effects compared with DML, but in general, their results are similar. Both the parametric and DML estimates suggest an important mediating role for household income. According to the DML estimates, for example, the interventional indirect effect operating through household income is $-.047$ standard deviations, which accounts for roughly 40% of the overall effect of education. These findings broadly align with our parametric estimates of interventional effects reported earlier in Chapter 4. This alignment suggests that our inferences about the role of income in transmitting the effect of college attendance on depression are robust to model misspecification.

6.4 Robust Estimation of Path-specific Effects

In this section, we outline robust approaches to estimating path-specific effects (PSEs) in analyses of multiple mediators. Specifically, we introduce two MR estimators for PSEs that can be viewed as extensions of $\hat{\psi}^{mr_1}_{d^*,d}$ and $\hat{\psi}^{mr_2}_{d^*,d}$ from Section 6.2, which targeted natural direct and indirect effects (Miles et al. 2020; Zhou 2022). For simplicity, we focus on PSEs involving two causally ordered mediators. A generalization of these methods for applications with $K(\geq 2)$ mediators is available in Zhou (2022).

With two causally ordered mediators, denoted by M_1 and M_2, the average total effect of an exposure D on an outcome Y can be decomposed into three different PSEs, as outlined in Chapter 5. Specifically, if we let $\psi_{d_1,d_2,d} = \mathbb{E}\left[Y\left(d, M_1\left(d_1\right), M_2\left(d_2, M_1\left(d_1\right)\right)\right)\right]$, then the average total effect can be decomposed as follows:

$$ATE\left(d, d^*\right) = \underbrace{\psi_{d^*,d^*,d} - \psi_{d^*,d^*,d^*}}_{PSE_{D\rightarrow Y}(d,d^*)} + \underbrace{\psi_{d^*,d,d} - \psi_{d^*,d^*,d}}_{PSE_{D\rightarrow M_2\rightarrow Y}(d,d^*)} + \underbrace{\psi_{d,d,d} - \psi_{d^*,d,d}}_{PSE_{D\rightarrow M_1\rightsquigarrow Y}(d,d^*)}, \tag{6.26}$$

where $PSE_{D\rightarrow Y}\left(d, d^*\right)$ represents the direct effect of the exposure on the outcome that does not operate through either mediator, $PSE_{D\rightarrow M_2\rightarrow Y}\left(d, d^*\right)$ represents an effect of the exposure that operates through M_2 but not M_1, and $PSE_{D\rightarrow M_1\rightsquigarrow Y}\left(d, d^*\right)$ represents an effect of the exposure transmitted via M_1, some of which may operate through the influence of M_1 on M_2.

To identify and estimate these PSEs, it suffices to identify and estimate $\psi_{d_1,d_2,d}$ for any combination of d_1, d_2, and d. Under assumptions (f.i) to (f.iii) from Chapter 5, $\psi_{d_1,d_2,d}$ is identified with the following expression:

$$\begin{aligned}\psi_{d_1,d_2,d} &= \sum_{c,m_1,m_2} \mathbb{E}\left[Y|c, d, m_1, m_2\right] P\left(m_2|c, d_2, m_1\right) P\left(m_1|c, d_1\right) P\left(c\right), \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y|C, d, M_1, M_2\right]\middle|C, d_2, M_1\right]\middle|C, d_1\right]\right] \end{aligned} \tag{6.27}$$

which is known as generalized mediation functional (GMF; Zhou 2022). This identification formula can also be expressed as follows:

$$\begin{aligned}
\psi_{d_1,d_2,d} = \mathbb{E}\Bigg[& \underbrace{\frac{I(D=d)}{P(d|C)}\left(\frac{P(M_2|C,d_2,M_1)\,P(M_1|C,d_1)}{P(M_2|C,d,M_1)\,P(M_1|C,d)}\right)(Y-\mathbb{E}[Y|C,d,M_1,M_2])}_{\text{term 1}} \\
& + \underbrace{\frac{I(D=d_2)}{P(d_2|C)}\left(\frac{P(M_1|C,d_1)}{P(M_1|C,d_2)}\right)\left(\mathbb{E}[Y|C,d,M_1,M_2]-\sum_{m_2}\mathbb{E}[Y|C,d,M_1,m_2]\,P(m_2|C,d_2,M_1)\right)}_{\text{term 2}} \\
& + \underbrace{\begin{array}{l}\displaystyle\frac{I(D=d_1)}{P(d_1|C)}\Bigg(\sum_{m_2}\mathbb{E}[Y|C,d,M_1,m_2]\,P(m_2|C,d_2,M_1) \\ \displaystyle\qquad\qquad -\sum_{m_1,m_2}\mathbb{E}[Y|C,d,M_1,m_2]\,P(m_2|C,d_2,M_1)\,P(m_1|C,d_1)\Bigg)\end{array}}_{\text{term 3}} \\
& + \underbrace{\sum_{m_1,m_2}\mathbb{E}[Y|C,d,M_1,m_2]\,P(m_2|C,d_2,M_1)\,P(m_1|C,d_1)}_{\text{term 4}}\Bigg],
\end{aligned} \tag{6.28}$$

where the fourth term in this expression is equivalent to $\mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[Y|C,d,M_1,M_2]\,\big|C,d_2,M_1\big]\,\Big|C,d_1\Big]$. Thus, Equation 6.28 differs from Equation 6.27 in that it contains three additional terms that each have a mean of zero–term 1, term 2, and term 3.

This alternative version of the identification formula suggests an estimator for $\psi_{d_1,d_2,d}$ with the following form:

$$\hat{\psi}^{mr_1}_{d_1,d_2,d} = \frac{1}{n}\sum \hat{S}^{mr_1}_{d_1,d_2,d}, \tag{6.29}$$

$$\begin{aligned}
\text{where}\quad \hat{S}^{mr_1}_{d_1,d_2,d} = & \frac{I(D=d)}{\hat{\pi}_d(C)}\left(\frac{\hat{P}(M_2|C,d_2,M_1)\,\hat{P}(M_1|C,d_1)}{\hat{P}(M_2|C,d,M_1)\,\hat{P}(M_1|C,d)}\right)(Y-\hat{\mu}_d(C,M_1,M_2)) \\
& + \frac{I(D=d_2)}{\hat{\pi}_{d_2}(C)}\left(\frac{\hat{P}(M_1|C,d_1)}{\hat{P}(M_1|C,d_2)}\right)\left(\hat{\mu}_d(C,M_1,M_2)-\sum_{m_2}\hat{\mu}_d(C,M_1,m_2)\,\hat{P}(m_2|C,d_2,M_1)\right) \\
& + \frac{I(D=d_1)}{\hat{\pi}_{d_1}(C)}\Bigg(\sum_{m_2}\hat{\mu}_d(C,M_1,m_2)\,\hat{P}(m_2|C,d_2,M_1) \\
& \qquad\qquad -\sum_{m_1,m_2}\hat{\mu}_d(C,M_1,m_2)\,\hat{P}(m_2|C,d_2,M_1)\,\hat{P}(m_1|C,d_1)\Bigg) \\
& + \sum_{m_1,m_2}\hat{\mu}_d(C,M_1,m_2)\,\hat{P}(m_2|C,d_2,M_1)\,\hat{P}(m_1|C,d_1).
\end{aligned} \tag{6.30}$$

In this expression, $\pi_D(C)$ denotes the conditional probability of the exposure D given the baseline confounders C, $P(M_1|C,D)$ denotes the conditional probability of the first mediator M_1 given the exposure D and baseline confounders C, $P(M_2|C,D,M_1)$ denotes the conditional probability of the second mediator M_2 given the exposure D, baseline confounders C, and the first mediator M_1, and lastly, $\mu_D(C,M_1,M_2)$ denotes the conditional mean of the outcome Y under exposure D given the baseline confounders C and both mediators M_1 and M_2. The "hats" over these terms denote that they are sample estimates.

Constructing $\hat{\psi}^{mr_1}_{d_1,d_2,d}$ requires four different models, which constitute the estimator's nuisance functions. These include an exposure model for $\pi_D(C)$, an outcome model for $\mu_D(C,M_1,M_2)$, and two mediator models for $P\left(M_1|C,D\right)$ and $P\left(M_2|C,D,M_1\right)$. The estimator is multiply robust in that it remains consistent for $\psi_{d_1,d_2,d}$ under assumptions (f.i) to (f.iii), even when any one of these four models is misspecified. In other words, it is consistent when models for $\pi_D\left(C\right)$, $P\left(M_1|C,D\right)$, and $P\left(M_2|C,D,M_1\right)$ are correctly specified; when models for $\pi_D\left(C\right)$, $P\left(M_1|C,D\right)$, and $\mu_D\left(C,M_1,M_2\right)$ are correctly specified; when models for $\pi_D\left(C\right)$, $P\left(M_2|C,D,M_1\right)$, and $\mu_D\left(C,M_1,M_2\right)$ are correctly specified; or when models for $P\left(M_1|C,D\right)$, $P\left(M_2|C,D,M_1\right)$, and $\mu_D\left(C,M_1,M_2\right)$ are correctly specified. Additionally, $\hat{\psi}^{mr_1}_{d_1,d_2,d}$ will be $\sqrt{n}$-consistent and asymptotically normal if the four nuisance functions are all estimated using parametric models and at least three of them are correctly specified. If all of these models are correctly specified, then $\hat{\psi}^{mr_1}_{d_1,d_2,d}$ will also be nonparametrically efficient.

This MR estimator necessitates fitting models for the conditional probability of each mediator, which is relatively easy to accomplish when both mediators are univariate and discrete with only a few values. However, when either mediator is multivariate, continuous, or discrete with many values, $\hat{\psi}^{mr_1}_{d_1,d_2,d}$ is much more difficult to implement.

In analyses of PSEs involving multivariate, continuous, or many-valued mediators, the ratios of mediator probabilities in term 1 and term 2 of Equation 6.12 can be recast as odds ratios that only involve a set of exposure probabilities. Specifically, using Bayes' rule, we can translate the mediator probability ratios as follows:

$$\frac{P\left(M_2|C,d_2,M_1\right)P\left(M_1|C,d_1\right)}{P\left(M_2|C,d,M_1\right)P\left(M_1|C,d\right)} = \frac{P\left(d_2|C,M_1,M_2\right)P\left(d_1|C,M_1\right)P\left(d|C\right)}{P\left(d|C,M_1,M_2\right)P\left(d_2|C,M_1\right)P\left(d_1|C\right)}$$
$$\frac{P\left(M_1|C,d_1\right)}{P\left(M_1|C,d_2\right)} = \frac{P\left(d_1|C,M_1\right)P\left(d_2|C\right)}{P\left(d_2|C,M_1\right)P\left(d_1|C\right)}$$

In addition, we can also recast the sums in term 2 and term 3 of Equation 6.12 as iterated expectations, such that the identification formula for $\psi_{d_1,d_2,d}$ can be alternatively expressed as follows:

$$\begin{aligned}\psi_{d_1,d_2,d} = \mathbb{E}\Bigg[& \underbrace{\frac{I\left(D=d\right)}{P\left(d_1|C\right)}\left(\frac{P\left(d_2|C,M_1,M_2\right)P\left(d_1|C,M_1\right)}{P\left(d|C,M_1,M_2\right)P\left(d_2|C,M_1\right)}\right)\left(Y-\mathbb{E}\left[Y|C,d,M_1,M_2\right]\right)}_{\text{term 1}} \\ & +\underbrace{\frac{I\left(D=d_2\right)}{P\left(d_1|C\right)}\left(\frac{P\left(d_1|C,M_1\right)}{P\left(d_2|C,M_1\right)}\right)\left(\mathbb{E}\left[Y|C,d,M_1,M_2\right]-\mathbb{E}\left[\mathbb{E}\left[Y|C,d,M_1,M_2\right]\middle|C,d_2,M_1\right]\right)}_{\text{term 2}} \\ & +\underbrace{\frac{I\left(D=d_1\right)}{P\left(d_1|C\right)}\left(\mathbb{E}\left[\mathbb{E}\left[Y|C,d,M_1,M_2\right]\middle|C,d_2,M_1\right]-\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y|C,d,M_1,M_2\right]\middle|C,d_2,M_1\right]\middle|C,d_1\right]\right)}_{\text{term 3}} \\ & +\underbrace{\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y|C,d,M_1,M_2\right]\middle|C,d_2,M_1\right]\middle|C,d_1\right]}_{\text{term 4}}\Bigg]. \end{aligned} \tag{6.31}$$

Equation 6.31 differs from Equation 6.28 in that it no longer contains any terms involving the conditional probability of either mediator.

This version of the identification formula for $\psi_{d_1,d_2,d}$ suggests an alternative estimator, which can be expressed as follows:

$$\hat{\psi}^{mr_2}_{d_1,d_2,d} = \frac{1}{n}\sum \hat{S}^{mr_2}_{d_1,d_2,d}, \tag{6.32}$$

$$\begin{aligned}\text{where}\quad \hat{S}^{mr_2}_{d_1,d_2,d} =& \frac{I\left(D=d\right)}{\hat{\pi}_{d_1}\left(C\right)}\left(\frac{\hat{\pi}_{d_2}\left(C,M_1,M_2\right)\hat{\pi}_{d_1}\left(C,M_1\right)}{\hat{\pi}_{d}\left(C,M_1,M_2\right)\hat{\pi}_{d_2}\left(C,M_1\right)}\right)\left(Y-\hat{\mu}_d\left(C,M_1,M_2\right)\right)\\ &+\frac{I\left(D=d_2\right)}{\hat{\pi}_{d_1}\left(C\right)}\left(\frac{\hat{\pi}_{d_1}\left(C,M_1\right)}{\hat{\pi}_{d_2}\left(C,M_1\right)}\right)\left(\hat{\mu}_d\left(C,M_1,M_2\right)-\hat{\nu}_{d_2}\left(C,M_1\right)\right)\\ &+\frac{I\left(D=d_1\right)}{\hat{\pi}_{d_1}\left(C\right)}\left(\hat{\nu}_{d_2}\left(C,M_1\right)-\hat{\xi}_{d_1}\left(C\right)\right)\\ &+\hat{\xi}_{d_1}\left(C\right).\end{aligned}\tag{6.33}$$

In Equation 6.33, $\pi_D\left(C\right)$ and $\mu_D\left(C,M_1,M_2\right)$ are defined exactly as before. In addition, $\pi_D\left(C,M_1,M_2\right)$ denotes the conditional probability of exposure D given the baseline confounders C and both mediators M_1 and M_2, while $\pi_D\left(C,M_1\right)$ denotes the conditional probability of the exposure given the baseline confounders and the first mediator only. Finally, $\nu_D\left(C,M_1\right)$ denotes the conditional mean of $\mu_d\left(C,M_1,M_2\right)$ under exposure D given the baseline confounders C and the first mediator M_1, while $\xi_D\left(C\right)$ denotes the conditional mean of $\nu_{d_2}\left(C,M_1\right)$ under exposure D given the baseline confounders C. Thus, the estimator $\hat{\psi}^{mr_2}_{d_1,d_2,d}$ requires fitting six different models: three exposure models for $\pi_D\left(C\right)$, $\pi_D\left(C,M_1\right)$, and $\pi_D\left(C,M_1,M_2\right)$, and three outcome models for $\mu_D\left(C,M_1,M_2\right)$, $\nu_D\left(C,M_1\right)$, and $\xi_D\left(C\right)$. These constitute its nuisance functions. Under assumptions (f.i) to (f.iii), the estimator is multiply robust in that it remains consistent as long as at least one of the following conditions is met: the models for $\pi_D\left(C\right)$, $\pi_D\left(C,M_1\right)$, and $\pi_D\left(C,M_1,M_2\right)$ are correctly specified; the models for $\pi_D\left(C\right)$, $\pi_D\left(C,M_1\right)$, and $\mu_D\left(C,M_1,M_2\right)$ are correctly specified; the models for $\pi_D\left(C\right)$, $\mu_D\left(C,M_1,M_2\right)$, and $\nu_D\left(C,M_1\right)$ are correctly specified; or the models for $\mu_D\left(C,M_1,M_2\right)$, $\nu_D\left(C,M_1\right)$, and $\xi_D\left(C\right)$ are correctly specified. Although $\hat{\psi}^{mr_2}_{d_1,d_2,d}$ involves two more nuisance functions than $\hat{\psi}^{mr_1}_{d_1,d_2,d}$, it does not involve models for the mediators. It is therefore much easier to use when any of the mediators are multivariate, continuous, or discrete but with many values. Similar to $\hat{\psi}^{mr_1}_{d_1,d_2,d}$, the estimator $\hat{\psi}^{mr_2}_{d_1,d_2,d}$ will be $\sqrt{n}$-consistent and asymptotically normal if its nuisance functions are estimated using parametric models and at least one of the four conditions outlined previously is met. It will also be nonparametrically efficient if the models for all six of its nuisance functions are correctly specified.

In practice, multiply robust estimates of the path-specific effects defined in Equation 6.26 can also be obtained from multiply robust estimates of the natural direct and indirect effects with respect to M_1 and the multivariate natural direct effect with respect to $\mathbf{M} = (M_1, M_2)$. Specifically, multiply robust estimates of the path-specific effects $PSE_{D\to Y}\left(d,d^*\right)$, $PSE_{D\to M_2\to Y}\left(d,d^*\right)$, and $PSE_{D\to M_1\rightsquigarrow Y}\left(d,d^*\right)$ can be constructed as follows:

$$\begin{aligned}\widehat{PSE}_{D\to Y}\left(d,d^*\right)^{mr_2} &= \widehat{NDE}_{\mathbf{M}}\left(d,d^*\right)^{mr_2}\\ \widehat{PSE}_{D\to M_2\to Y}\left(d,d^*\right)^{mr_2} &= \widehat{NDE}_{M_1}\left(d,d^*\right)^{mr_2} - \widehat{NDE}_{\mathbf{M}}\left(d,d^*\right)^{mr_2}\\ \widehat{PSE}_{D\to M_1\rightsquigarrow Y}\left(d,d^*\right)^{mr_2} &= \widehat{NIE}_{M_1}\left(d,d^*\right)^{mr_2},\end{aligned}\tag{6.34}$$

where $\widehat{NIE}_{M_1}\left(d,d^*\right)^{mr_2}$, $\widehat{NDE}_{M_1}\left(d,d^*\right)^{mr_2}$, and $\widehat{NDE}_{\mathbf{M}}\left(d,d^*\right)^{mr_2}$ are computed with the multiply robust estimator $\hat{\psi}^{mr_2}_{d^*,d}$ introduced in Section 6.2. These estimates of the PSEs will be equivalent to those obtained from $\hat{\psi}^{mr_2}_{d_1,d_2,d}$ directly.

Like all the other MR estimators we have considered in this chapter, the biases of $\hat{\psi}^{mr_1}_{d_1,d_2,d}$ and $\hat{\psi}^{mr_2}_{d_1,d_2,d}$

are both governed by a sum of bias products, which renders them amenable for use with DML. When implemented using DML, both MR estimators for $\psi_{d_1,d_2,d}$ will be $\sqrt{n}$-consistent, asymptotically normal, and nonparametrically efficient if assumptions (f.i) to (f.iii) are met and provided that the algorithms used to fit their nuisance functions converge at a rate faster than $n^{1/4}$. In this case, the sampling variance of $\hat{\psi}^{mr_1}_{d_1,d_2,d}$ can be estimated as follows:

$$\widehat{\mathrm{Var}}\left(\hat{\psi}^{mr_1}_{d_1,d_2,d}\right) = \frac{1}{n}\widehat{\mathrm{Var}}\left(\hat{S}^{mr_1}_{d_1,d_2,d}\right), \tag{6.35}$$

while the sampling variance for corresponding estimators of total and path-specific effects can be estimated with the following expressions:

$$\begin{aligned}
\widehat{\mathrm{Var}}\left(\widehat{PSE}_{D\to Y}\left(d,d^*\right)^{mr_1}\right) &= \frac{1}{n}\widehat{\mathrm{Var}}\left(\hat{S}^{mr_1}_{d^*,d^*,d} - \hat{S}^{mr_1}_{d^*,d^*,d^*}\right) \\
\widehat{\mathrm{Var}}\left(\widehat{PSE}_{D\to M_2\to Y}\left(d,d^*\right)^{mr_1}\right) &= \frac{1}{n}\widehat{\mathrm{Var}}\left(\hat{S}^{mr_1}_{d^*,d,d} - \hat{S}^{mr_1}_{d^*,d^*,d}\right) \\
\widehat{\mathrm{Var}}\left(\widehat{PSE}_{D\to M_1\rightsquigarrow Y}\left(d,d^*\right)^{mr_1}\right) &= \frac{1}{n}\widehat{\mathrm{Var}}\left(\hat{S}^{mr_1}_{d,d,d} - \hat{S}^{mr_1}_{d^*,d,d}\right) \\
\widehat{\mathrm{Var}}\left(\widehat{ATE}\left(d,d^*\right)^{mr_1}\right) &= \frac{1}{n}\widehat{\mathrm{Var}}\left(\hat{S}^{mr_1}_{d,d,d} - \hat{S}^{mr_1}_{d^*,d^*,d^*}\right).
\end{aligned} \tag{6.36}$$

Analogous formulas can be used to estimate the variance of $\hat{\psi}^{mr_2}_{d_1,d_2,d}$ and its corresponding estimators for the effects of interest, whenever these estimators are constructed via DML. The variance estimates obtained from Equations 6.35 and 6.36 can then be used to compute standard errors, confidence intervals, and p-values, all using standard inferential methods based on the normal distribution.

Alternatively, when $\hat{\psi}^{mr_1}_{d_1,d_2,d}$ and $\hat{\psi}^{mr_2}_{d_1,d_2,d}$ are implemented with parametric models for their nuisance functions, inferential statistics should be computed using the nonparametric bootstrap. In this situation, the variance formulas in Equations 6.35 and 6.36 are only valid if all the nuisance functions are correctly modeled, even though the estimators for $\psi_{d_1,d_2,d}$ are themselves robust to multiple forms of misspecification. Using the bootstrap for inference obviates this concern in parametric applications of MR estimation.

With data from the NLSY, we used MR estimation to reanalyze the effects of college attendance on depression transmitted through unemployment status (M_1) and household income (M_2). Following our analyses from previous chapters, we transformed household income into its natural logarithm and adjusted for the same set of baseline confounders as before. Since household income is a continuous variable, we estimated the total and path-specific effects of interest using $\hat{\psi}^{mr_2}_{d_1,d_2,d}$ only, which does not require distribution models for the mediators. To compute these estimates, we implemented $\hat{\psi}^{mr_2}_{d_1,d_2,d}$ with both parametric models and DML. The parametric estimates were obtained from linear models for the outcome (CES-D scores) and logit models for the exposure (college attendance), while the DML estimates were computed using repeated cross-fitting with $J = 5$ partitions and a super learner composed of the LASSO and random forests for each nuisance function.

Table 6.4 presents results from this analysis. Both the parametric and DML approach suggest a sizeable total effect of college attendance on CES-D scores, but a minimal role for unemployment in mediating this relationship, as the estimated $\mathrm{PS}E_{D\to M_1\rightsquigarrow Y}(1,0)$ is consistently close to zero. The estimated $\mathrm{PS}E_{D\to M_2\to Y}(1,0)$, which captures the effect of college attendance transmitted through household income alone, is larger. According to the DML estimates, for example, this PSE is $-.037$, or about 30% of the total effect. These findings are mostly consistent with our parametric estimates reported in Chapter 5, although differences in the relative size of estimates for the $\mathrm{PS}E_{D\to M_2\to Y}(1,0)$ suggest that any conclusions about the unique mediating role of income may be somewhat sensitive to our modeling choices. Links to the code

Table 6.4: Multiply Robust Estimates of Total and Path-specific Effects of College Attendance on CES-D Scores, as Mediated by Unemployment and Household Income, from the NLSY.

Estimand	MR (Parametric)	DML
$ATE\,(1,0)$	$-.144$ $[-.253,-.035]$	$-.121$ $[-.218,-.024]$
$PSE_{D\to Y}\,(1,0)$	$-.119$ $[-.255,.016]$	$-.081$ $[-.190,.027]$
$PSE_{D\to M_2\to Y}\,(1,0)$	$-.017$ $[-.081,.048]$	$-.037$ $[-.071,-.003]$
$PSE_{D\to M_1\rightsquigarrow Y}\,(1,0)$	$-.008$ $[-.026,.010]$	$-.002$ $[-.015,.011]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals. For the parametric estimates, the confidence intervals were computed using the nonparametric bootstrap with $B = 2000$ replications. For the DML estimates, confidence intervals are constructed using ±1.96 times the analytical standard errors described in Section 6.4. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch6/table_6-4`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/NLSY79`.

and data used for this analysis are available in the table footnote.

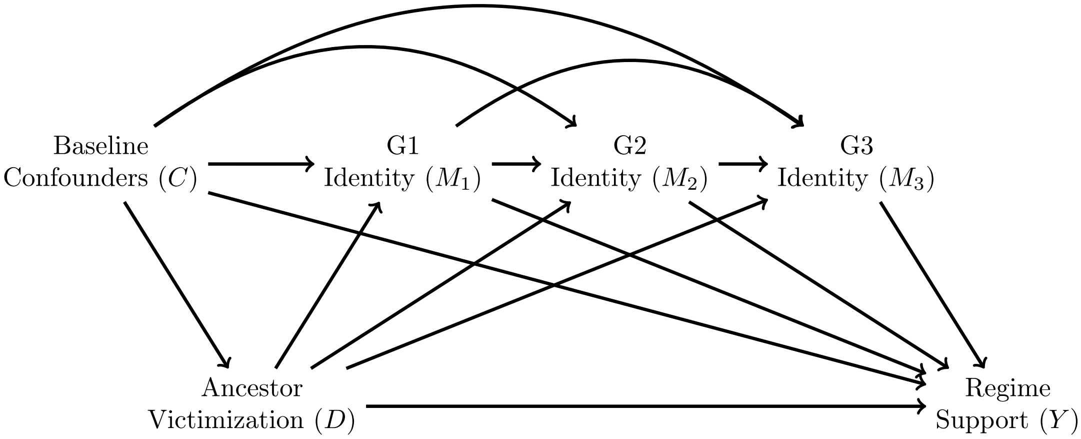
6.5 An Empirical Illustration: The Legacy of Political Violence

In this section, we illustrate MR estimation for PSEs through a reanalysis of data from Lupu and Peisakhin (2017). In 2014, these authors conducted a multigenerational survey of Crimean Tatars, a minority Muslim population living in Crimea, to study the legacy of political violence that occurred during the deportation of Crimean Tatars from their homeland to Central Asia in 1944. Due to starvation and infectious diseases, a sizable portion of the deportees died during or shortly after the deportation. Yet, "[a]lthough all Crimean Tatars suffered the violence of deportation, some lost more family members along the way" (pg. 837). Leveraging this variation in violent victimization, the authors found that the grandchildren of individuals who suffered a greater number of family deaths are more likely to support the Crimean Tatar political leadership, hold more hostile attitudes toward Russia, and participate more in politics.

To investigate the intergenerational pathways that transmit the legacy of political violence, the authors conducted an "implicit mediation analysis" by adding measures of descendants' political identity into their main regression models and assessing the changes in the coefficients on ancestor victimization. This approach is problematic, however, because descendants' political identities are likely shaped by the political identities of their parents and grandparents, which might also affect descendant political attitudes and behavior. In other words, the political identities of first- and second-generation respondents are potential exposure-induced confounders for the mediator-outcome relationship–that is, for the relationship between descendants' identities and their political attitudes–implying that the natural indirect effect via descendants' identities cannot be nonparametrically identified.

To circumvent this identification problem, we analyze the political identities of first-, second-, and third-generation respondents as three causally ordered mediators, following Zhou and Yamamoto (2022). We then focus on estimating the total and path-specific effects of ancestor victimization on respondent attitudes toward Russia's annexation of Crimea. The hypothesized causal relationships in these data can be represented using the DAG in Figure 6.1. In this DAG, the exposure D represents ancestor victimization, a binary

Figure 6.1: Causal Pathways from Ancestor Victimization to Descendants' Regime Support.



variable denoting whether a family member of the first-generation respondent died during or shortly after the 1944 deportation. The political identities of first-, second-, and third-generation respondents constitute the mediators of interest in this analysis, which are denoted by M_1, M_2, and M_3 respectively. They are measured by the intensity of a respondent's attachment to the Crimean Tatars as a social group, their association of that group with victimhood, and their perception of the threat posed by Russia. The outcome, denoted by Y, captures whether the third-generation respondent supported Russia's annexation of Crimea. Finally, the baseline confounders C include measures of the first-generation respondent's family wealth, religiosity, attitudes toward the Soviet Union, and experience with persecution by state authorities prior to deportation. These variables are used to control for potential confounding of the exposure-mediator, exposure-outcome, and mediator-outcome relationships.

With this notation, the PSEs of ancestor victimization on regime support can be expressed as follows:

$$\begin{aligned}
PSE_{D\to Y}\left(1,0\right) &= NDE_{M_1,M_2,M_3}\left(1,0\right) \\
PSE_{D\to M_3\to Y}\left(1,0\right) &= NDE_{M_1,M_2,M_3}\left(1,0\right) - NDE_{M_1,M_2}\left(1,0\right) \\
PSE_{D\to M_2\rightsquigarrow Y}\left(1,0\right) &= NDE_{M_1,M_2}\left(1,0\right) - NDE_{M_1}\left(1,0\right) \\
PSE_{D\to M_1\rightsquigarrow Y}\left(1,0\right) &= NIE_{M_1}\left(1,0\right).
\end{aligned}$$

The $PSE_{D\to Y}(1,0)$ represents a direct effect of ancestor victimization that does not operate through the political identities of first-, second-, or third-generation respondents. The $PSE_{D\to M_3\to Y}(1,0)$ represents the effect of victimization that operates solely through the political identities of third-generation respondents along the path $D \to M_3 \to Y$. The $PSE_{D\to M_2\rightsquigarrow Y}(1,0)$ represents the effect that operates through the political identities of second-generation, but not first-generation, respondents along the paths $D \to M_2 \to Y$ and $D \to M_2 \to M_3 \to Y$. Finally, the $PSE_{D\to M_1\rightsquigarrow Y}(1,0)$ represents the effect of victimization that operates through the political identities of first-generation respondents, including any influence of these identities on the identities of subsequent generations. In other words, it captures the effects transmitted along the paths $D \to M_1 \to Y$, $D \to M_1 \to M_2 \to Y$, $D \to M_1 \to M_3 \to Y$, and $D \to M_1 \to M_2 \to M_3 \to Y$. It is therefore equivalent to the natural indirect effect of victimization with respect to M_1.

We estimate these effects using a generalization of the robust methods described in Section 6.4 for

Table 6.5: Multiply Robust Estimates for the Total and Path-Specific Effects of Ancestor Victimization on Regime Support as Mediated by Political Identities.

Estimand	Parametric MR	DML
$ATE(1,0)$	$-.246$ $[-.364, -.128]$	$-.208$ $[-.284, -.132]$
$PSE_{D\to Y}(1,0)$	$-.109$ $[-.199, -.020]$	$-.093$ $[-.153, -.034]$
$PSE_{D\to M_3\to Y}(1,0)$	$-.033$ $[-.083, .017]$	$-.008$ $[-.042, .026]$
$PSE_{D\to M_2\rightsquigarrow Y}(1,0)$	$-.036$ $[-.090, .018]$	$-.027$ $[-.057, .002]$
$PSE_{D\to M_1\rightsquigarrow Y}(1,0)$	$-.067$ $[-.121, -.013]$	$-.080$ $[-.118, -.042]$

Note: Estimates are expressed in standard deviation units. The numbers in parentheses represent 95% confidence intervals. For the parametric estimates, the confidence intervals were computed using the nonparametric bootstrap with $B = 2000$ replications. For the DML estimates, confidence intervals are constructed using ±1.96 times the analytical standard errors described in Section 6.4. The code used to produce these results is available at `https://github.com/causalMedAnalysis/repFiles/tree/main/code/ch6/table_6-5`, and the data are available at `https://github.com/causalMedAnalysis/repFiles/tree/main/data/Tatar`.

the PSEs of three causally ordered mediators (Zhou 2022). Since the three mediators in this analysis are multivariate, where each is composed of separate measures for attachment, association, and threat perception, we focus on the second MR estimator, $\hat{\psi}^{mr_2}_{d_1,d_2,d}$, which only requires models for the exposure and outcome. We implemented this estimator using both parametric models and DML. The parametric estimates were obtained from logit models for both the outcome and the exposure, while the DML estimates were based on repeated cross-fitting with $J = 5$ partitions and a super learner for each nuisance function composed of the LASSO and a random forest.

Table 6.5 presents the results from this analysis. In general, the parametric and DML estimates are very similar. Consistent with Lupu and Peisakhin (2017), we find that ancestor victimization significantly reduces support for Russia's annexation of Crimea among third-generation descendants. According to the DML estimates, for example, victimization reduces regime support by about 21 percentage points (i.e., $-.208$ on the probability scale). Both the parametric and DML estimates also suggest that more than half of the total effect is indirect, operating through the political identities of first-, second-, and third-generation respondents. In particular, the political identities of grandparents appear to play an especially important explanatory role. The estimated $PSE_{D\to M_1\rightsquigarrow Y}(1,0)$ is -0.067 according to the parametric MR approach and -0.08 according to DML. By contrast, the estimated PSEs through the political identities of second- and third-generation respondents, $PSE_{D\to M_2\rightsquigarrow Y}(1,0)$ and $PSE_{D\to M_3\to Y}(1,0)$, are all relatively small, and their confidence intervals span zero. This suggests that exposure to political violence affects the identities of first-generation respondents, who then pass their identities down through the family line, shaping the attitudes of their descendants several generations later.

6.6 Summary

In this chapter, we introduced methods for analyzing causal mediation that exhibit several optimal characteristics. Specifically, these methods are multiply robust, efficient, and asymptotically normal under fairly general conditions. They are multiply robust in that they remain consistent for their target parameters, assuming they are nonparametrically identified, even under multiple forms of model misspecification. The efficiency of these methods is evident in their ability to produce estimates with the smallest sampling variance achievable within a nonparametric model for the data. We discussed and illustrated several such methods throughout the chapter, including a doubly robust estimator for the average total effect, two estimators for natural effects that are triply robust, and several multiply robust estimators for interventional and path-specific effects.

In addition to efficiency and multiple robustness, another advantage of the estimators discussed in this chapter is their compatibility with machine learning methods. We specifically highlighted how these estimators can be implemented with debiased machine learning (DML), an approach that employs data-adaptive algorithms with a sample-splitting procedure known as repeated cross-fitting to estimate the requisite nuisance functions. When DML is used and the effects of interest are nonparametrically identified, the resulting estimator will be $\sqrt{n}$-consistent, asymptotically normal, and nonparametrically efficient as long as the machine learning estimates for the nuisance functions converge at a sufficiently fast rate. The DML approach therefore offers even stronger protection against bias due to model misspecification, without much loss of efficiency.

Despite the many advantages of robust estimation methods and DML, they are not without limitations. In particular, all the methods discussed in this chapter require discrete exposures for effective implementation, and they generally perform best with binary, polytomous, or ordinal exposures that have only a limited number of categories. For continuous exposures, the estimators outlined previously are not appropriate, and even with exposures that are discrete but have many values, these methods can be difficult to implement, and their performance may deteriorate in practice. While there have been several advances in robust estimation for the total effects of continuous exposures (Diaz and van der Laan 2013; Kennedy et al. 2017), similar techniques are not yet available for analyses of causal mediation.

To limit our focus and simplify the presentation of complex material, this chapter omitted certain advanced topics surrounding robust estimation methods for causal mediation analysis. These include robust estimation of natural direct and indirect effects in the presence of unobserved confounding (Dukes et al. 2023), robust estimation of controlled direct effects (Zhou 2020), robust estimation of interventional effects in the presence of multivariate or continuous exposure-induced confounders (Diaz et al. 2021), and a detailed exposition of non- and semi-parametric efficiency theory with application to causal inference (Hines et al. 2022; Kennedy 2016; Tsiatis 2006). We refer researchers interested in these topics to the specialized literature cited above.

Chapter 7

Mediation Analysis with (Quasi-)Experimental Methods

In this book, we have focused mainly on methods of causal mediation analysis that require strong assumptions about the absence of several different forms of confounding. For example, in Chapters 4 to 6, we used data from the 1979 National Longitudinal Survey of Youth (NLSY; Bureau of Labor Statistics 2019) to investigate whether income mediates the effect of college attendance on mental health. Our analyses throughout these chapters assumed complete measurement and proper control of all relevant confounders. Assumptions about unobserved confounding were necessary because the NLSY is an observational study, lacking any experimental control over the exposure or mediator of interest. In these data, individual behaviors and societal structures combined to determine the educational attainment and subsequent income levels of participants, rather than some external intervention or manipulation. Despite these methodological constraints, our results suggested that college attendance mitigates depression partly through its influence on earnings, but the validity of these findings remains open to uncertainty and debate, due to their reliance on strong, unverifiable assumptions.

This uncertainty stems largely from the omnipresent threat of bias due to unobserved confounding. In our analysis of the NLSY, many unobserved variables could confound the effects of interest, including socioemotional skills, disability status, or childhood parenting practices, among a variety of other possibilities. All of these factors may jointly influence educational attainment, income, and mental health, leading to systematic errors in our estimates of total, direct, and indirect effects. Although we made every effort to measure and adjust for an exhaustive set of confounding variables, our conclusions based on these analyses must remain cautious and provisional, and lingering questions persist. Is the impact of college attendance on mental health mediated by income, or do unobserved factors that influence these variables drive the observed relationships?

In analyses of causal effects, researchers often use randomized experiments to control for potential confounders (Rubin 1974; Hernan and Robins 2020; Morgan and Winship 2014). A standard experimental design involves randomly assigning participants to different levels of the exposure, which, in expectation, ensures that all potential confounders are evenly distributed across exposure groups. Those assigned to different exposures can then be compared in order to cleanly estimate causal effects, as any difference in outcomes between these groups can be confidently attributed to the exposure itself rather than other extraneous variables. While highly effective at eliminating certain types of bias, a standard experimental design does not completely resolve the problem of confounding in analyses of causal mediation. Indeed, we have encountered these challenges in several of our prior illustrations. For example, we previously reanalyzed data from the JOBSII experiment in Chapter 3 (Vinokur et al. 1995; Vinokur and Schul 1997; Imai et al. 2010a), which was designed to investigate how a job training workshop impacts employment and whether various psychological factors might transmit this effect. Even with the benefit of random assignment to the job training workshop in this study, our mediation analysis still depended on strong, unverifiable assumptions about the absence of confounding variables that were not satisfied by the experimental design. These assumptions were less stringent than those required of purely observational studies, like the NLSY, but they justified nontrivial concerns about the validity of our results nonetheless.

Can we design studies that are capable of identifying and estimating direct and indirect effects without relying on strong, often unrealistic assumptions about the absence of confounding? In this chapter, we introduce experimental methods tailored specifically for analyzing causal mediation. We begin by reviewing why standard experiments, which manipulate only the exposure, still require unverifiable assumptions to address the problem of confounding bias in mediation analyses. Next, we explore several alternative designs that incorporate random assignment of mediators, in addition to the exposure. These include joint and sequential randomization designs (Hernan and Robins 2020), a parallel randomization design (Acharya et al. 2018; Imai et al. 2013), and multi-arm randomization (Moreno-Betancur and Carlin 2018). We explain how these types of experiments can obviate concerns about multiple forms of unobserved confounding in analyses of direct and indirect effects, thereby reducing the risk of bias and bolstering the credibility of inferences about causal mediation.

After outlining a series of experimental designs, we then extend our discussion to quasi-experimental approaches, including parallel encouragement designs, instrumental variable models, and difference-in-difference analyses (Imai et al. 2013; Burgess et al. 2015). Despite the promise of randomized experiments designed to identify direct and indirect effects, experimental mediation analysis remains “harder than it looks” (Bullock and Ha 2011). Among other challenges, experimental analyses of mediation are complicated by the logistical complexity and practical difficulty of directly manipulating both an exposure and mediator. In cases where direct manipulation is infeasible, researchers might resort to quasi-experiments, which exploit natural sources of random variation or indirect manipulations of the exposure and mediator to analyze causal mechanisms. Although these designs do not afford the same degree of experimental control over the exposure or mediator, they are capable of generating estimates for direct and indirect effects under alternative and, sometimes, more defensible assumptions than those required in observational studies.

Nevertheless, while quasi-experimental methods may provide a more accessible and defensible alternative for social science applications, they are not without their own limitations. In particular, moving from an observational to a quasi-experimental design often involves supplanting one set of unverifiable assumptions with another that can be just as contentious. Mediation analysis is exceptionally challenging, and no singular approach can completely circumvent difficult and potentially problematic assumptions. It is still essential, however, to incorporate experimental and quasi-experimental techniques into the methodological toolkit for investigating causal mechanisms, despite their limitations. By employing a variety of approaches, each predicated on different assumptions of varying strength, researchers can attenuate uncertainty and bolster the credibility of inferences. Ultimately, the most convincing conclusions about causal mediation will come from triangulating results across observational, experimental, and quasi-experimental designs, grounding inferences in a broad spectrum of evidence.

Throughout this chapter, we draw on a range of experimental studies from across the social sciences to illustrate key concepts and methods. As experimental and quasi-experimental techniques for analyzing causal mediation are still nascent, few established examples of their application presently exist. Thus, our illustrations predominantly revolve around hypothetical modifications of several recent experiments, initially conducted with different aims, to enable a more robust analysis of mediation. We offer these examples as templates that researchers can adapt for implementing the proposed designs in the future.

7.1 Conventional Experiments and Their Limitations

In standard experiments, only the exposure of interest is randomly assigned. Following this manipulation, the outcome and any putative mediators are then measured for each participant. While randomization of the exposure alone is sufficient to identify total effects, it does not help to uncover the causal processes that link the exposure to the outcome.

To illustrate, let's revisit the JOBSII study (Vinokur et al. 1995; Vinokur and Schul 1997; Imai et al. 2010a), which we previously discussed in Chapter 3. This study investigated the impact of a job training workshop on subsequent employment among individuals looking for work. Participants were randomly assigned to either a workshop designed to improve their job search skills or a control group that received only a booklet with very basic advice for finding a new job. The study sought to determine whether the workshop boosted employment by increasing participants' confidence in their ability to find new work, a construct the researcher's referred to as "job search self-efficacy."

Figure 7.1 displays a directed acyclic graph (DAG) that outlines a set of causal relationships among variables in the JOBSII experiment, where only the exposure was randomized. In this graph, D represents the exposure, encoded as a binary variable where 1 indicates assignment to the job training workshop and 0 indicates assignment to the control group. The outcome, denoted by Y, is another binary variable indicating whether a participant had secured at least part-time employment by the study's conclusion. The mediator, denoted by M, is a measure of job search self-efficacy, constructed from survey questions that assessed participants' confidence in their abilities to complete an application, locate job openings through their social networks, and so on. The figure additionally contains four other variables, denoted by X, V, Z, and L, which represent unobserved factors that may influence the mediator and outcome in different ways. And, it demonstrates how randomization in a standard experiment effectively severs all causal links from potential confounding variables to the exposure, as depicted by the dotted, faded arrows from X and V into D.

Because randomization ensures that there is not any unobserved confounding of the exposure-outcome relationship, a key assumption necessary for identifying the average total effect of D on Y is met by design in a standard experiment. Specifically, random assignment of the exposure D guarantees its independence from all other determinants of the outcome, whether they are observed or not. As a result, potentially confounding variables like X or V cannot contaminate any association observed between the exposure and outcome. In this situation, a straightforward comparison of sample means of the outcome in the treatment and control groups provides an unbiased, consistent, and nonparametric estimate of the average total effect. Additionally adjusting for baseline covariates, as in Chapter 3, can improve the precision of estimates (i.e., attenuate random variability due to sampling), but it is not essential for satisfying key identification assumptions about the absence of unobserved confounding, which are met by the design of the experiment.

While standard experiments cleanly identify total effects, they fall short of satisfying the conditions for identifying controlled, natural, and interventional effects, which variously capture how the impact of exposure on the outcome is mediated. Consider, for example, the controlled direct effect of D on Y, setting $M = m$. Identifying this effect hinges on the absence of unobserved confounding for both the exposure-

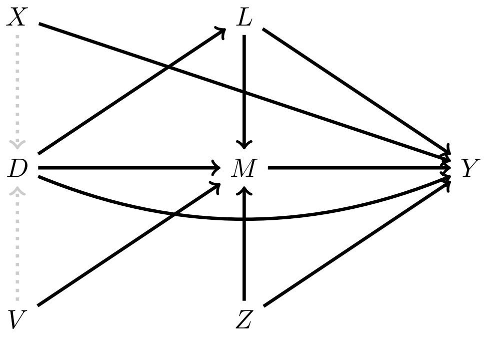


Figure 7.1: A Graphical Mediation Model Depicting Causal Relations in a Standard Experimental Design with Random Assignment of the Exposure.

Note: D denotes the exposure, M denotes a focal mediator, and Y denotes the outcome. X, V, Z and L are unobserved variables that may affect the mediator or outcome in different ways. The dotted, faded arrows denote causal paths that are broken via randomization.

outcome and mediator-outcome relationships. However, in a standard experimental design, randomization only guarantees that the relationship between the exposure and outcome is unconfounded. This limitation is depicted in Figure 7.1, where the exposure D is independent of all potential confounders, but the effect of M on Y is confounded by Z and L because the mediator was not randomly assigned. Although identifying and consistently estimating the controlled direct effect would remain possible if Z and L were observed, as outlined in Chapter 4, accounting for all such confounders in practice is challenging. Thus, analyses of controlled direct effects in standard experiments still depend on an unverifiable assumption about the absence of mediator-outcome confounding that is often dubious and can undermine the credibility of inferences about causal mediation.

The standard experimental design presents similar challenges for identifying interventional direct and indirect effects. To identify the interventional effects of D on Y through M, we must eliminate any confounding not only between the exposure and outcome, and the mediator and outcome, but also between the exposure and mediator. Random assignment in a standard experiment eliminates exposure-outcome and exposure-mediator confounding, which is indicated by the dotted, faded arrows from both X and V to D in Figure 7.1. Nevertheless, as discussed previously for controlled direct effects, randomizing the exposure does not obviate the problem of mediator-outcome confounding. In standard experiments, then, identifying and consistently estimating interventional effects also requires measuring and properly adjusting for all variables that jointly affect the mediator and outcome. This is a formidable task, given the difficulty of ensuring that every relevant factor has been accounted for.

Standard experiments fall short of satisfying the assumptions needed to identify natural direct and indirect effects as well. As outlined in Chapter 3, identifying the natural effects of D on Y through M is contingent upon the absence of unobserved confounding for the exposure-outcome, exposure-mediator, and mediator-outcome relationships. Additionally, it also requires that there not be any confounding of the mediator-outcome relationship that is exposure-induced, whether by observed or unobserved variables. While random assignment in a standard experiment ensures that the exposure-outcome and exposure-mediator relationships are unconfounded, it does not eliminate the potential for mediator-outcome confounding by either unobserved or exposure-induced factors. This limitation is depicted by the causal relations involving Z and L in Figure 7.1, which are undisturbed by randomizing the exposure D alone. As a result, analyses of natural effects in standard experiments also depend on strong assumptions about the absence of several different forms of mediator-outcome confounding that can be difficult to justify in practice.

The standard experimental design is a powerful tool. It enables researchers to draw credible inferences about total effects. However, it does not permit an uncomplicated assessment of the causal mechanisms hypothesized to explain them (Heckman and Smith 1995; Cook 2002). Despite these challenges, data from standard experiments are frequently used to investigate causal mechanisms, and mediation analyses in such applications are conducted under assumptions that are neither verifiable nor met by the experimental design itself (e.g., Brader et al. 2008; Vinokur et al. 1995; Vinokur and Schul 1997). This does not imply that standard experiments should never be used to analyze mediation. Compared to purely observational studies, mediation analyses conducted with data from a standard experiment are generally more robust (Bullock and Ha 2011; Green et al. 2010), since these designs satisfy at least some of the necessary conditions for identifying controlled, interventional, and natural effects. Nevertheless, randomizing the exposure alone is not the only experimental approach available for analyzing mediation. In the next section, we outline alternative experimental designs tailored specifically for identifying causal mechanisms.

7.2 Experimental Designs for Analyzing Causal Mediation

In this section, we introduce a series of experimental designs that can illuminate causal mechanisms without relying on stringent assumptions about the absence of unobserved confounding. Each design strategically employs random manipulations of *both* the exposure and mediator. We first discuss a single-arm experiment, where the exposure and mediator are either jointly or sequentially randomized, which can be used to identify controlled direct effects. Next, we discuss a parallel design that combines an experiment based on joint or sequential randomization with a standard experiment in which only the exposure is randomly assigned. This design can identify total and controlled direct effects together. Finally, we describe a multi-arm experiment capable of identifying interventional direct and indirect effects. Along the way, we also explain how the parallel and multi-arm designs can facilitate identification of natural effects as well, whenever the exposure and mediator do not interact to influence the outcome.

We illustrate these designs by adapting and extending a resume correspondence experiment aimed at uncovering racial discrimination against Black job applicants in the labor market. In a resume correspondence experiment, researchers distribute fabricated resumes to employers advertising entry-level job openings. These resumes are identical, except for the name given to the fictitious applicant. For example, each resume might indicate that the applicant recently completed a college degree and has a consistent work history, but some feature a Black-sounding name like “Lakisha Washington” or “Jamal Jones,” while others bear a White-sounding name, such as “Emily Walsh” or “Greg Baker,” strategically chosen to manipulate the perceived racial identity of the applicant (Bertrand and Mullainathan 2004). The identical resumes, distinguished solely by the names attached to them, are randomly distributed to employers, and subsequent callbacks for interviews are closely monitored by the researchers. Under this design, the total effect of race on callback rates, among applicants with the credentials specified on the fabricated resume, can be identified and consistently estimated, similar to any standard experiment.

Many studies employing variants of this basic design have found that fictitious applicants with Black names are significantly less likely to receive callbacks than those with White names (Bertrand and Mullainathan 2004; Kline et al. 2022; Quillian and Midtboen 2021; Quillian and Lee 2023). But what drives this disparity? While correspondence experiments accurately quantify the total amount of racial discrimination at the point of interviewing, they do not shed much light on the possible mechanisms through which employer perceptions of race may influence their decision-making. One possible explanation for the lower rate of callbacks among Black applicants could be that employers indulge a "taste for discrimination" (Becker 2010; Krueger 1963). According to this hypothesis, some employers may simply avoid Black applicants altogether, irrespective of their qualifications, skills, or ability to perform the job well. Alternatively, the "racial proxy" hypothesis posits that employers may rely on an applicant's race as a crude substitute for other information that is more relevant to judgments about their potential job performance but is not immediately available from the job application (Heckman 1998; Neumark and Rich 2019). Despite resumes presenting identical levels of education and work experience, employers might use the race of the applicant to speculate about unlisted qualifications or capabilities. For example, in an experiment where all resumes indicate that the applicant has a university degree, racial stereotypes could still lead employers to assume that Black applicants did not perform as well academically in college compared to their White counterparts. Thus, the preference for White applicants may not stem from an outright distaste for Black workers but rather from a flawed, stereotypical inference about their academic achievements in college.

In this situation, researchers might aim to understand if, and to what degree, an applicant's race affects callbacks because employers use it to make assumptions about their academic performance. For example, they could explore whether including additional details on the resume, like cumulative grade point average (GPA), might neutralize the effect of race on callbacks. They could also examine how much discrimination might be eliminated by virtue of providing this additional information, comparing the overall impact of race with its effect when GPA is disclosed. This investigation could even explore how race leads to stereotypical judgments about academic performance, and how these judgments influence the prospects of an interview in turn. As we outline in the sections below, a standard correspondence design can be adapted and extended to study these causal mechanisms, shedding light on whether discrimination stems from immovable biases or from the misuse of race as a proxy for academic skills.

7.2.1 Joint and Sequential Randomization Designs

All necessary assumptions for nonparametrically identifying and consistently estimating controlled direct effects can be satisfied by design in experiments that employ joint or sequential randomization of the exposure and a focal mediator. In experiments using a joint randomization design, participants are randomly assigned to different levels of both the exposure and the mediator at baseline. In a sequential design, participants are initially assigned at random to different exposures; then, following randomization of the exposure, they are randomly assigned again to different levels of the mediator within each exposure group.

Figure 7.2 displays a DAG that describes causal relationships among variables in an experiment where the exposure and mediator are jointly or sequentially randomized. As before, D represents the exposure, while the mediator and outcome are denoted by M and Y, respectively. The figure illustrates how joint or sequential randomization severs all causal paths from potential confounding variables to both the exposure and mediator. This is depicted by the dotted, faded arrows from X and V to D and from V, L, and Z to M, which represent paths that are "broken," or eliminated, via randomization.

Because random assignment in a joint or sequential design ensures that the exposure-outcome and the mediator-outcome relationships are both unconfounded, it satisfies the key assumptions for identifying controlled direct effects, defined as $CDE\,(d, d^{*}, m) = \mathbb{E}\,[Y\,(d, m) - Y\,(d^{*}, m)]$. In particular, randomization of the exposure D and mediator M guarantees their mutual independence from other determinants of the outcome Y. As a result, variables like X, V, L or Z cannot engender confounding bias in this setting, whether

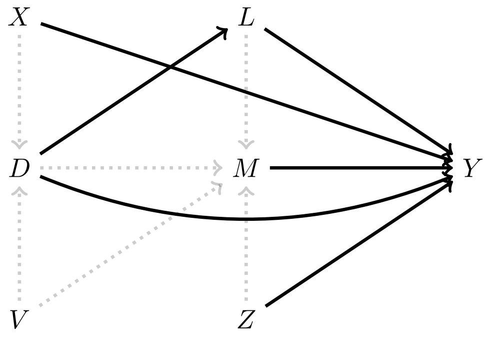


Figure 7.2: A Graphical Mediation Model Depicting Causal Relations in Joint or Sequential Experiments with Random Assignment of both the Exposure and Mediator.

Note: D denotes the exposure, M denotes a focal mediator, and Y denotes the outcome. X, V, Z and L are unobserved variables that may affect the outcome in different ways. The dotted, faded arrows denote causal paths that are broken via randomization.

they are observed or not. In both joint and sequentially randomized experiments, then, simply comparing sample means of the outcome among participants assigned to different levels of the exposure, but the same level of the mediator, will provide unbiased and consistent estimates for controlled direct effects.

Specifically, a nonparametric estimator for controlled direct effects in an experiment using a joint or sequential design can be expressed as follows:

$$\begin{aligned}\widehat{CDE}\left(d, d^{*}, m\right)^{jsx} &= \hat{\mathbb{E}}\left[Y|d, m\right] - \hat{\mathbb{E}}\left[Y|d^{*}, m\right] \\ &= \bar{Y}_{d,m} - \bar{Y}_{d^{*},m}, \end{aligned} \tag{7.1}$$

where $\bar{Y}_{d,m}$ and $\bar{Y}_{d^{*},m}$ represent sample means of the outcome for participants assigned to $D = d$ and $M = m$, and to $D = d^{*}$ and $M = m$, respectively. This expression is a variant of the nonparametric estimator discussed in Chapter 3, which additionally included adjustments for a set of baseline covariates denoted by C. However, in joint or sequential experiments, these adjustments are not necessary to identify and consistently estimate controlled direct effects, as randomization of the exposure and mediator effectively eliminates any confounding by these factors. Therefore, they need not be incorporated in Equation 7.1, and the "jsx" superscript denotes that this expression is a nonparametric estimator specifically for a joint or sequential experiment in which confounder adjustment is unnecessary.

Nevertheless, even in a joint or sequential experiment, adjusting for baseline covariates might improve the precision of effect estimates. This improvement can be achieved using any of the parametric methods described in Chapter 3 for estimating controlled direct effects. By applying these methods to data from a joint or sequential experiment that also include baseline variables measured before random assignment, researchers may be able to attenuate sampling variability in their estimates.

To illustrate, let's reconsider the resume correspondence experiment designed to detect racial discrimination against Black job applicants. In a standard version of this experiment, only the applicant's name—which suggests their racial identity—is randomly manipulated on the fictitious resumes distributed to employers. Suppose, however, the researchers were now interested in whether including additional information about the applicants, such as their cumulative GPA in college, might mitigate the effect of race on the likelihood of receiving a callback. In this example, the exposure D is coded 1 for applicants with a Black-sounding name and 0 for those with a White-sounding name. The mediator M represents the applicant's cumulative GPA, measured on a four-point scale. And the outcome Y is a binary variable, coded 1 to indicate that an application received a callback from an employer and 0 otherwise. The target estimand is a controlled direct effect, $CDE\left(1,0,m\right)=\mathbb{E}\left[Y\left(1,m\right)-Y\left(0,m\right)\right]$, which captures the difference in callback rates for fictitious applicants with Black- versus White-sounding names when they are indicated to have the same GPA m.

To identify and estimate this effect, the researchers could adapt a standard correspondence experiment by including an additional line on the applicant's resume that reports their GPA. They could then randomly assign this value along with the name of the applicant, following a joint randomization design. For example, they could use two distinct values for the putative mediator to represent different levels of achievement in college: a high-performing applicant with a 4.0 GPA and an applicant with average performance, signaled by a 3.0 GPA. The modified resumes, now differentiated by Black- versus White-sounding names and by high versus low GPAs, could then be randomly distributed to employers and the resulting callbacks recorded.

Under this experimental design, the controlled direct effect of race on callbacks, accounting for the applicant's GPA in college, can be nonparametrically identified and consistently estimated. Specifically, the controlled direct effect of race for applicants with high achievement could be estimated as $\widehat{CDE}\left(1,0,4\right)^{jsx}=\bar{Y}_{1,4}-\bar{Y}_{0,4}$, where $\bar{Y}_{1,4}$ is the observed callback rate for applicants with a Black-sounding name ($D=1$) and a high GPA ($M=4.0$), and $\bar{Y}_{0,4}$ is the callback rate for applicants with a White-sounding name ($D=0$) and the same high GPA. Similarly, an estimate of the other controlled direct effect for those with lower achievement could also be computed simply by comparing callback rates among applicants with a 3.0 GPA and Black- versus White-sounding names.

Confidence intervals and hypothesis tests could be constructed using the nonparametric bootstrap, as detailed in Chapter 3, or by using standard parametric methods (Campbell and Stanley 2015; Montgomery 2019), if appropriate. Should these inferential statistics indicate that the controlled direct effects are either substantively negligible or statistically indistinguishable from zero, the researchers might conclude that disclosing an applicant's GPA on their resume effectively counteracts racial discrimination in the labor market. Conversely, if the effect estimates are substantial and significantly different from zero, it would suggest that including information about an applicant's college GPA does not eliminate racial discrimination at the point of selecting job candidates to interview.

7.2.2 A Parallel Randomization Design

In a parallel randomization design, participants are initially assigned at random to one of two experimental arms. Then, in the first arm, only the exposure is randomized, mirroring a standard experiment. In the second arm, by contrast, both the exposure and mediator are randomized together, as in the joint randomization design described previously. The combination of standard and joint randomization across two separate arms of a single experiment allows parallel randomization designs to satisfy all the assumptions necessary for nonparametrically identifying and consistently estimating both total and controlled direct effects (Pearl 2001). In addition, under the assumption that the exposure and mediator do not interact to influence the outcome, natural direct and indirect effects can also be identified from an experiment adopting this approach (Imai et al. 2013).

Figure 7.3 presents a DAG that describes causal relations among variables in an experiment with parallel randomization. Panel A illustrates these relations in the first arm of the experiment, where only the exposure D is randomized. This panel shows how randomization in a standard experiment eliminates all causal links from potential confounding variables to the exposure. Similarly, panel B illustrates causal relations among

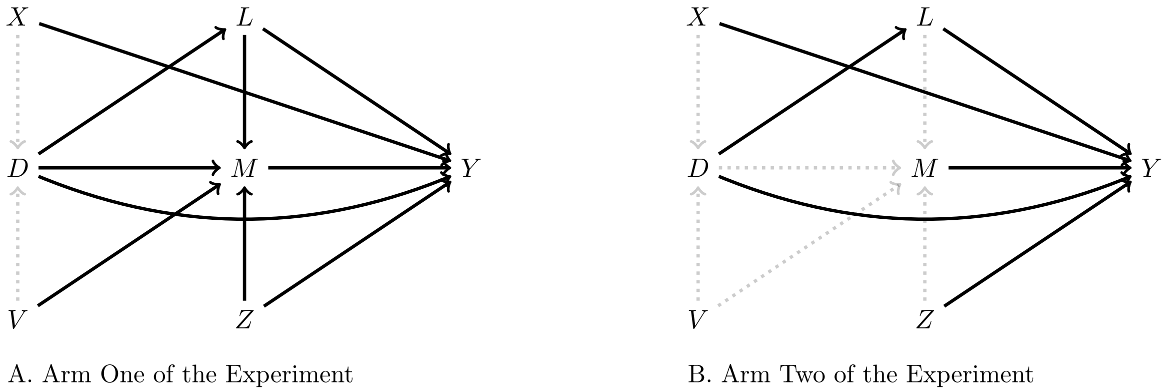


Figure 7.3: Graphical Mediation Models Depicting Causal Relations in a Parallel Experiment with Random Assignment of the Exposure in Arm One and Random Assignment of both the Exposure and Mediator in Arm Two.

Note: D denotes the exposure, M denotes a focal mediator, and Y denotes the outcome. X, V, Z and L are unobserved variables that may affect the mediator and/or the outcome in different ways. The dotted, faded arrows denote causal paths that are broken via randomization.

variables in the second arm of the experiment, where both the exposure and the mediator are randomly assigned. It shows how joint randomization severs all causal paths from potential confounders to both the exposure D and mediator M.

Random assignment in the first arm of a parallel design satisfies all the conditions necessary for nonparametrically identifying the average total effect of the exposure D on the outcome Y, which can be formally defined as $ATE\left(d, d^*\right) = \mathbb{E}\left[Y\left(d\right) - Y\left(d^*\right)\right]$. This effect can be consistently estimated simply by comparing sample means of the outcome among participants assigned to different levels of the exposure in the first arm of the experiment. Specifically, let A denote the experimental arm to which a participant is assigned, with 1 indicating the first arm and 2 the second arm. In addition, let $\bar{Y}_{d,a}$ represent the sample mean of the outcome for participants assigned to exposure $D = d$ in arm $A = a$. With this notation, a nonparametric estimator for the average total effect is given by the following expression:

$$\begin{aligned} \widehat{ATE}\left(d, d^*\right)^{px} &= \hat{\mathbb{E}}\left[Y|d, A = 1\right] - \hat{\mathbb{E}}\left[Y|d^*, A = 1\right] \\ &= \bar{Y}_{d,1} - \bar{Y}_{d^*,1}, \end{aligned} \tag{7.2}$$

where the superscript "px" denotes that this estimator is specific to the parallel randomization design.

Similarly, random assignment in the second arm of the experiment satisfies all the conditions necessary for nonparametrically identifying the controlled direct effect of the exposure D on the outcome Y, with the mediator M set to the same value m for everyone. This effect is formally defined as $CDE\left(d, d^*, m\right) = \mathbb{E}\left[Y\left(d, m\right) - Y\left(d^*, m\right)\right]$, and it can be consistently estimated by comparing sample means of the outcome among participants assigned to different levels of the exposure, but the same level of the mediator, in the second arm of the experiment. Specifically, let $\bar{Y}_{d,m,a}$ denote the sample mean of the outcome for participants assigned to exposure $D = d$ and mediator $M = m$ in arm $A = a$. A nonparametric estimator for the controlled direct effect can then be expressed as follows:

$$\begin{aligned}\widehat{CDE}\left(d, d^{*}, m\right)^{px} &= \hat{\mathbb{E}}\left[Y|d, m, A=2\right] - \hat{\mathbb{E}}\left[Y|d^{*}, m, A=2\right] \\ &= \bar{Y}_{d,m,2} - \bar{Y}_{d^{*},m,2}, \end{aligned} \tag{7.3}$$

where $A = 2$ denotes assignment to the second arm and the "px" superscript again indicates that this estimator is tailored for a parallel randomization design.

Parallel randomization also enables nonparametric identification and consistent estimation of the difference between the $ATE\left(d, d^{*}\right)$ and $CDE\left(d, d^{*}, m\right)$, known as the *average eliminated effect* (Acharya et al. 2016; Glynn 2021). This effect can be formally defined as follows:

$$AEE\left(d, d^{*}, m\right) = ATE\left(d, d^{*}\right) - CDE\left(d, d^{*}, m\right). \tag{7.4}$$

It quantifies the change in the impact of exposure D on the outcome Y following an intervention that controls the mediator at the same level m for everyone. The eliminated effect can be consistently estimated using the following expression with data from a parallel experiment:

$$\begin{aligned}\widehat{AEE}\left(d, d^{*}, m\right)^{px} &= \widehat{ATE}\left(d, d^{*}\right)^{px} - \widehat{CDE}\left(d, d^{*}, m\right)^{px} \\ &= \bar{Y}_{d,1} - \bar{Y}_{d^{*},1} - \bar{Y}_{d,m,2} + \bar{Y}_{d^{*},m,2}, \end{aligned} \tag{7.5}$$

where all terms are defined as before. Together, the total, controlled direct, and eliminated effects comprise the estimands that can be nonparametrically identified and consistently estimated under parallel randomization, without the need for additional assumptions that are not met by the design of the experiment.

Beyond these estimands, researchers are often interested in natural direct and indirect effects as well, which can be formally defined as $NDE\left(d, d^{*}\right) = \mathbb{E}\left[Y\left(d, M\left(d^{*}\right)\right) - Y\left(d^{*}, M\left(d^{*}\right)\right)\right]$ and $NIE\left(d, d^{*}\right) = \mathbb{E}\left[Y\left(d, M\left(d\right)\right) - Y\left(d, M\left(d^{*}\right)\right)\right]$, respectively. In general, there is no experimental design that can identify natural direct and indirect effects without additional assumptions. These effects cannot be identified from a parallel randomization design without additional assumptions because the $NDE\left(d, d^{*}\right)$ is not, in general, equivalent to the $CDE\left(d, d^{*}, m\right)$, nor does the $NIE\left(d, d^{*}\right)$ correspond with the $AEE\left(d, d^{*}, m\right)$. However, under certain conditions, these effects do align. Specifically, when the exposure and mediator do not interact to influence the outcome, the controlled and natural direct effects–as well as the eliminated and natural indirect effects–are equivalent. Thus, while a parallel experiment can only identify and consistently estimate total, controlled direct, and eliminated effects without additional assumptions, it can also recover natural effects if there is no exposure-mediator interaction (Imai et al. 2013).

The assumption of no exposure-mediator interaction is stringent and cannot be empirically verified. This is because the assumption must hold for every individual, not merely on average for the target population as a whole. Although data from a parallel experiment can be used to identify and estimate the expected value of any individual interaction effects–for example, by comparing $\widehat{CDE}\left(d, d^{*}, m\right)^{px}$ across values of m–it cannot recover the individual effects themselves. In this situation, the average of these effects might be zero even if all the individual effects are not. That is, the interaction effect for each individual in the target population may vary around an average of zero, which would still breach the assumption of no interaction in this context. Consequently, the assumption is not directly testable, and if the exposure and mediator do interact to influence the outcome for any individual, the parallel design cannot identify or consistently estimate natural effects. Under parallel randomization, then, analyzing natural effects no longer hinges on assumptions about the absence of unobserved or exposure-induced confounding, as outlined in Chapter 3. Instead, it requires an unverifiable assumption about the absence of exposure-mediator interaction at the individual level.

When this assumption holds, natural direct and indirect effects can be consistently estimated using data from a parallel randomization design. In this case, the natural direct effect will coincide with the controlled direct effect evaluated at any value of the mediator. Let $\hat{\pi}_{m|a}$ denote the proportion of participants for whom $M = m$ in arm $A = a$ of the experiment. A pooled estimator for the natural direct effect can then be expressed as follows:

$$\begin{aligned} \widehat{NDE}\left(d, d^{*}\right)^{px} &= \sum_{m} \widehat{CDE}\left(d, d^{*}, m\right)^{px} \hat{P}\left(m|A = 2\right) \\ &= \sum_{m} \left(\bar{Y}_{d,m,2} - \bar{Y}_{d^{*},m,2}\right) \hat{\pi}_{m|2}. \end{aligned} \tag{7.6}$$

This expression estimates the natural direct effect by first computing differences in the mean of the outcome between participants assigned to different levels of the exposure, d versus d^{*}, but the same value m of the mediator, all in the second arm of the experiment. It evaluates these differences at every level of the mediator and then adds them together, weighting each by the proportion of participants assigned to a particular mediator value. In other words, it averages point estimates for the controlled direct effect at each value m over the distribution of the mediator in the second arm of the experiment.

Similarly, an estimator for the natural indirect effect is given by the following expression:

$$\begin{aligned} \widehat{NIE}\left(d, d^{*}\right)^{px} &= \widehat{ATE}\left(d, d^{*}\right)^{px} - \widehat{NDE}\left(d, d^{*}\right)^{px} \\ &= \bar{Y}_{d,1} - \bar{Y}_{d^{*},1} - \sum_{m} \left(\bar{Y}_{d,m,2} - \bar{Y}_{d^{*},m,2}\right) \hat{\pi}_{m|2}, \end{aligned} \tag{7.7}$$

which just computes the difference between the total effect estimate in the first arm of the experiment and the direct effect estimate from the second arm. These two estimators, $\widehat{NDE}\left(d, d^{*}\right)^{px}$ and $\widehat{NIE}\left(d, d^{*}\right)^{px}$, are only consistent when applied to data from a parallel randomization design if there is no exposure-mediator interaction. Otherwise, they may be biased and inconsistent.

To illustrate, consider again a resume correspondence experiment designed to study racial discrimination in the labor market. In this example, recall that the exposure D is coded 1 for fictitious applicants with a Black-sounding name and 0 for those with a White-sounding name, the mediator M represents the applicant's cumulative GPA on a four-point scale, and the outcome Y is a binary variable indicating whether an application receives an employer callback.

Suppose the researchers were now aiming to determine both the overall level of racial discrimination and how much of this discrimination could be attenuated by providing additional information about an applicant's academic performance in college. The target estimands corresponding to the new aims of this study include: (i) an average total effect, $ATE\left(1, 0\right) = \mathbb{E}\left[Y\left(1\right) - Y\left(0\right)\right]$, which measures the overall difference in callback rates between applicants with Black- versus White-sounding names; (ii) a controlled direct effect, $CDE\left(1, 0, m\right) = \mathbb{E}\left[Y\left(1, m\right) - Y\left(0, m\right)\right]$, which measures the racial difference in callback rates when the applicants have the same GPA m; and (iii) an eliminated effect, $AEE\left(d, d^{*}, m\right) = ATE\left(d, d^{*}\right) - CDE\left(d, d^{*}, m\right)$, which captures how the degree of discrimination changes after providing information about an applicant's GPA on their resume.

To identify and estimate these effects, the researchers could adopt a parallel randomization design with two distinct arms. The first arm would conduct a standard correspondence experiment, where only the name of the applicant is randomized, and information on their GPA is not included on the fictitious resumes. The second arm would conduct another experiment with a joint randomization design, as described in Section 7.2.1. In this arm, each applicant's resume would feature an additional line displaying their GPA. Along with their name, the applicant's GPA is also randomized to signal either high ($M = 4.0$) or average ($M = 3.0$) academic performance in college.

With this approach, the average total effect of race on callbacks could be consistently estimated as $\widehat{ATE}\left(1,0\right)^{px} = \bar{Y}_{1,1} - \bar{Y}_{0,1}$, where $\bar{Y}_{1,1}$ is the callback rate for applications with a Black-sounding name ($D = 1$) in the first arm of the experiment ($A = 1$), and $\bar{Y}_{0,1}$ represents the callback rate for those with a White-sounding name ($D = 0$) in the same arm. In addition, the controlled direct effect for applicants with high achievement could be consistently estimated as $\widehat{CDE}\left(1,0,4\right)^{px} = \bar{Y}_{1,4,2} - \bar{Y}_{0,4,2}$. In this expression,$\bar{Y}_{1,4,2}$ is the observed callback rate for applications with a Black-sounding name and a high GPA ($M = 4.0$), while $\bar{Y}_{0,4,2}$ is the callback rate for those with a White-sounding name and the same high GPA, both measured in the second arm of the experiment ($A = 2$). Finally, the eliminated effect, which quantifies the impact of signaling that an applicant has a high GPA on the level of discrimination, could be estimated by calculating the difference between $\widehat{ATE}\left(1,0\right)^{px}$ and $\widehat{CDE}\left(1,0,4\right)^{px}$.

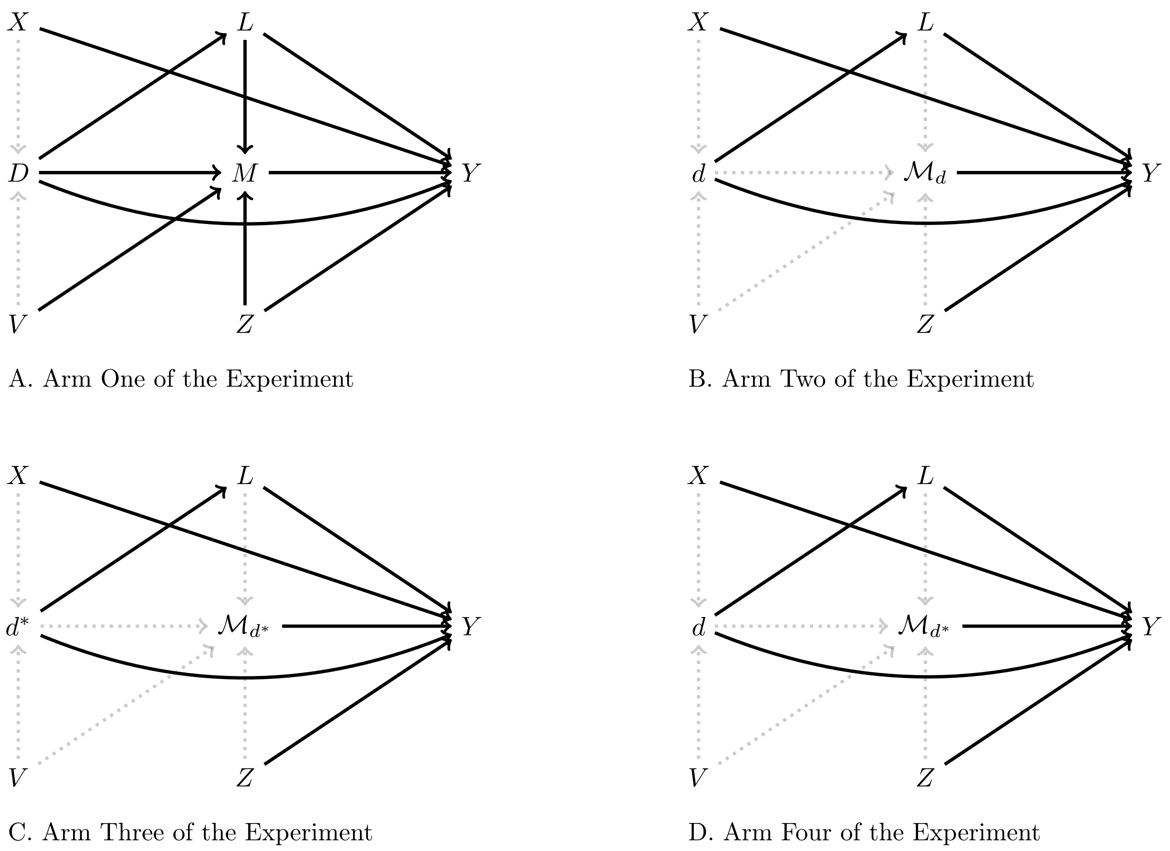
Assuming there is no interaction effect between race and GPA on the likelihood of a callback for all prospective employers, natural direct and indirect effects could also be consistently estimated from this experiment, as outlined in Equations 7.6 and 7.7. However, this assumption is not guaranteed to hold by design, and it is inconsistent with theory and prior scientific knowledge about interactions between race and other signals of job performance in employer decision making (Gaddis 2015, 2018; Pager 2003).

Inferential statistics for these effects could be constructed using the nonparametric bootstrap. If the eliminated effect is substantively small or its confidence interval spans zero, the researchers might conclude that disclosing an applicant's GPA on their resume does not appreciably reduce the degree of racial discrimination at the point of interviewing for a job, contrary to the racial proxy hypothesis.

7.2.3 A Multi-arm Randomization Design

A multi-arm randomization design can be used to nonparametrically identify and consistently estimate interventional direct and indirect effects of an exposure D on an outcome Y mediated by M. The interventional direct effect is defined as $IDE\left(d,d^*\right) = \mathbb{E}\left[Y\left(d,\mathcal{M}\left(d^*|C\right)\right) - Y\left(d^*,\mathcal{M}\left(d^*|C\right)\right)\right]$. In this expression, $\mathcal{M}\left(d|C\right)$ denotes a value of the mediator randomly selected from its distribution under level d of the exposure, given a set of baseline covariates C, which may be empty, and $\mathcal{M}\left(d^*|C\right)$ is defined analogously. Similarly, the interventional indirect effect is defined as $IIE\left(d,d^*\right) = \mathbb{E}\left[Y\left(d,\mathcal{M}\left(d|C\right)\right) - Y\left(d,\mathcal{M}\left(d^*|C\right)\right)\right]$, and the sum of these estimands equals an overall effect, which is given by $OE\left(d,d^*\right) = \mathbb{E}\left[Y\left(d,\mathcal{M}\left(d|C\right)\right) - Y\left(d^*,\mathcal{M}\left(d^*|C\right)\right)\right]$.

In a mutli-arm design targeting interventional effects, participants are initially randomized into one of four experimental arms, denoted by A. Then, in the first arm ($A = 1$), the exposure alone is randomly assigned, as in a standard experiment, and the mediator is measured for all participants in this arm. Let $\hat{f}(M|d, A = 1)$ denote the empirical distribution of the mediator among participants assigned to exposure d in arm one of the experiment, and let $\hat{f}(M|d^*, A = 1)$ represent the same distribution but now among those exposed to d^* in the first arm. In addition, let $\mathcal{M}_d$ denote a value selected at random from the distribution



A. Arm One of the Experiment

B. Arm Two of the Experiment

C. Arm Three of the Experiment

D. Arm Four of the Experiment

Figure 7.4: Graphical Mediation Models Depicting Causal Relations in a Multi-arm Experiment with Random Assignment of the Exposure in Arm One and Assignment to Different Values of the Exposure and Random Draws of the Mediator in Arms Two through Four.

Note: D denotes the exposure, M denotes a focal mediator, and Y denotes the outcome. X denotes a set of potential confounders affecting both the exposure and the outcome, V denotes a set of potential confouders for the relationship between the exposure and mediator, and Z denotes a set of mediator-outcome confounders. L denotes another set of mediator-outcome confounders that are exposure-induced. $\mathcal{M}_d$ denotes a value of the mediator randomly selected from its distribution among those exposed to d in the first arm of the experiment, while $\mathcal{M}_{d^*}$ represents a value of the mediator randomly selected from its distribution among those exposed to d^*, also in the first arm. The dotted, faded arrows denote causal paths that are broken via randomization.

$\hat{f}(M|d, A = 1)$, with $\mathcal{M}_{d^*}$ defined analogously.

In the second arm ($A = 2$), participants are assigned to exposure d and a value for the mediator, $\mathcal{M}_d$, randomly drawn from $\hat{f}(M|d, A = 1)$, its distribution among those exposed to d in the first arm of the experiment. In the third arm ($A = 3$), participants are assigned to exposure d^* and a value for the mediator, $\mathcal{M}_{d^*}$, randomly drawn from $\hat{f}(M|d^*, A = 1)$, its distribution among those exposed to d^* in the first arm. Finally, in the fourth arm ($A = 4$), participants are assigned to exposure d and a value for the mediator, $\mathcal{M}_{d^*}$, selected randomly from $\hat{f}(M|d^*, A = 1)$, its distribution among those exposed to d^* in the first arm.

With this design, the first arm of the experiment recovers the distribution of the mediator under different levels of the exposure. This distribution is then used in subsequent arms of the experiment to assign values of the mediator at random, enabling identification and consistent estimation for the interventional effects of interest.

Figure 7.4 presents a DAG that describes causal relations among variables in a multi-arm experiment. Panel A shows these relations in the first arm of the experiment, where only the exposure is randomized, eliminating any exposure-outcome or exposure-mediator confounding. Panels B through D depict these relations in the second, third, and fourth arms of the experiment, where random assignment of participants to different arms also eliminates any exposure-outcome or exposure-mediator confounding. Moreover, random assignment of the mediator within these arms additionally eliminates any mediator-outcome confounding. Comparing sample means of the outcome among participants assigned to the second, third, and fourth arms the experiment therefore provides unbiased and consistent estimates for interventional direct and indirect effects.

Specifically, let $\bar{Y}_a$ denote the sample mean of the outcome among participants assigned to arm $A = a$ of the experiment. A nonparametric estimator for the interventional direct effect in a multi-arm design can then be expressed as follows:

$$\begin{aligned} \widehat{IDE}\left(d, d^*\right)^{mx} &= \hat{\mathbb{E}}\left[Y|A=4\right] - \hat{\mathbb{E}}\left[Y|A=3\right] \\ &= \bar{Y}_4 - \bar{Y}_3. \end{aligned} \tag{7.8}$$

In this expression, $\bar{Y}_4$ represents the sample mean of the outcome in arm $A = 4$, where each participant is assigned to level d of the exposure and a value of the mediator, given by $\mathcal{M}_{d^*}$, randomly drawn from the distribution observed among those exposed to d^* in the first arm of the experiment. Similarly, $\bar{Y}_3$ is the sample mean of the outcome in arm $A = 3$, where participants are assigned to level d^* of the exposure and a value of the mediator, $\mathcal{M}_{d^*}$, selected at random from its distribution among those also exposed to d^* in the first arm.

A nonparametric estimator for the interventional indirect effect can be expressed as follows:

$$\begin{aligned} \widehat{IIE}\left(d, d^*\right)^{mx} &= \hat{\mathbb{E}}\left[Y|A=2\right] - \hat{\mathbb{E}}\left[Y|A=4\right] \\ &= \bar{Y}_2 - \bar{Y}_4. \end{aligned} \tag{7.9}$$

In this expression, $\bar{Y}_4$ is the sample mean of the outcome in arm $A = 4$, as outlined previously, while $\bar{Y}_2$ represents the sample mean of the outcome in arm $A = 2$. In this arm, participants are assigned to level d of the exposure and a value of the mediator, $\mathcal{M}_d$, randomly drawn from the distribution among those also exposed to d in arm one.

A nonparametric estimator for the overall effect can be obtained by summing the estimators for the interventional direct and indirect effects, which yields the following expression:

$$\begin{aligned} \widehat{OE}\left(d, d^*\right)^{mx} &= \widehat{IDE}\left(d, d^*\right)^{mx} + \widehat{IIE}\left(d, d^*\right)^{mx} \\ &= \bar{Y}_2 - \bar{Y}_3. \end{aligned} \tag{7.10}$$

Across all these estimators, the “mx” superscript indicates that they are designed specifically for an experiment with a multi-arm design.

To summarize, then, the interventional direct effect is estimated by comparing the mean of the outcome in the fourth arm with the third arm. The interventional indirect effect is estimated by comparing the outcome mean in the second arm with the fourth arm. And the overall effect is estimated by comparing the the outcome mean in the second arm with the third arm.

Multi-arm experiments of this type cannot nonparametrically identify and consistently estimate natural direct and indirect effects, except under special circumstances. As with a parallel randomization design, natural direct and indirect effects can be identified and estimated in a multi-arm experiment when the exposure and mediator do not interact to influence the outcome. This condition is neither guaranteed nor empirically verifiable in a multi-arm experiment, and it should only be invoked cautiously. However, in cases where there is indeed no exposure-mediator interaction, interventional direct and indirect effects are equivalent to natural direct and indirect effects. Consequently, natural effects can then be estimated using the expressions provided in Equations 7.8 and 7.9 as well.

A resume correspondence experiment, designed to investigate racial discrimination, could also be adapted to estimate interventional direct and indirect effects. Consider a scenario where the researchers aim to evaluate the causal process hypothesized by the racial proxy perspective, where a job applicant's race may lead to stereotypical judgments about their academic performance among employers, which then impacts their prospects for an interview. In this example, the exposure D is coded 1 for applicants with Black-sounding names and 0 for those with White-sounding names. The mediator M represents the applicant's GPA on a four-point scale, and the outcome Y indicates whether an application receives a callback for an interview.

To investigate this causal process, the researchers could compile a sampling frame of employers advertising entry-level job openings and then randomly assign them to one of four experimental arms. The first arm would resemble a standard correspondence experiment, where fictitious resumes are distributed to employers without any information about GPAs and only the name of applicant is randomized. After distributing these resumes, the researchers would conduct follow-up interviews with employers, asking them to estimate the fictitious applicant's GPA on a four-point scale based only on the information available from their resume. The distribution of GPAs reported by employers in this arm would then be used to implement the subsequent experiments.

Specifically, after completing data collection in the first arm, the researchers would proceed with the experiments in arms two through four. In the second arm, fictitious resumes with Black-sounding names $(D = 1)$ would be distributed to employers, including an additional line displaying the applicant's GPA. The GPA displayed would be a random draw from the distribution of values reported by employers in the first arm who received a resume with a Black-sounding name, denoted by $\mathcal{M}_1$. In the third arm, the resumes would include White-sounding names $(D = 0)$ and a GPA randomly drawn from the distribution observed among employers in the first arm who received a resume with a White-sounding name, denoted by $\mathcal{M}_0$. Lastly, in the fourth arm, resumes with Black-sounding names $(D = 1)$ would include a GPA randomly drawn from the distribution observed among employers in the first arm who received a resume with a White-sounding name, again denoted by $\mathcal{M}_0$. After distributing all these resumes, the researchers would record which fictitious applicants receive callbacks for interviews.

The target estimands in this analysis include the interventional direct effect, $IDE\,(1,0) = \mathbb{E}\left[Y\left(1,\mathcal{M}\left(0\right)\right) - Y\left(0,\mathcal{M}\left(0\right)\right)\right]$, and the interventional indirect effect, $IIE\,(1,0) = \mathbb{E}\left[Y\left(1,\mathcal{M}\left(1\right)\right) - Y\left(1,\mathcal{M}\left(0\right)\right)\right]$, where the set of baseline covariates C is empty. In substantive terms, the interventional direct effect measures the racial disparity in callback rates that is not due to stereotypical perceptions among employers about racial differences in GPA. Conversely, the indirect effect specifically captures the impact of these stereotypical perceptions on the likelihood of receiving a callback. The

$IDE\,(1,0)$ would be estimated by comparing callback rates between the third and fourth arms of the experiment, while the $IIE\,(1,0)$ would be estimated by comparing callback rates in the second and fourth arms, as outlined in Equations 7.8 and 7.9.

Inferential statistics for these effects could be constructed using the nonparametric bootstrap. If the interventional indirect effect were substantively small or its confidence interval included zero, the researchers might conclude that employer perceptions of academic performance in college do not mediate the effect of race on the likelihood of selecting an applicant for a callback. These results would contradict the racial proxy hypothesis, which suggests that racial discrimination in the labor market occurs because employers use an applicant's race to infer other attributes related to their potential job performance, such as their prior academic performance.

7.2.4 Limitations

While experimental methods can facilitate identification and consistent estimation in analyses of causal mediation, they suffer from a number of practical limitations. One of the main challenges is the difficulty associated with directly manipulating a mediator (Bullock and Ha 2011; Imai et al. 2013). This challenge is especially pronounced in fields like psychology, sociology, and political science, where mediators of interest often involve complex constructs like emotions, attitudes, or cognition. These constructs are exceedingly difficult, if not impossible, to randomly assign or control at particular levels, which can complicate or even preclude experimental analyses of causal mediation in many applications.

Even when it is possible to experimentally manipulate a focal mediator, implementing a joint, parallel, or multi-arm design will often require considerable funding and a high degree of logistical coordination. Standard experiments in the social sciences are already costly and can be challenging to conduct. With experiments designed to analyze mediation, incorporating manipulations to additional variables, possibly across multiple arms of a study, only increases the costs and logistical demands of implementation, thereby limiting the feasibility of these designs in practice. In general, there exists a trade-off between the complexity and practicality of an experimental design and the strength of the assumptions needed to assess causal mediation (Imai et al. 2013).

Additionally, while experiments offer improved identification power for controlled, eliminated, and interventional effects, identifying natural effects still depends on strong, unverifiable assumptions. These assumptions concern not the absence of unobserved confounding, but rather the absence of interaction effects between the exposure and mediator on the outcome. Similar to the assumptions regarding unobserved confounding required in observational studies, the assumption of no exposure-mediator interaction in experimental studies aimed at identifying natural effects can be highly restrictive and often implausible. Thus, when analyzing natural effects, transitioning from an observational to a parallel or multi-arm experimental study merely substitutes one set of questionable assumptions for another.

Despite these limitations, experimental methods can still offer robust evidence about causal mechanisms in certain applications. Researchers interested in causal mediation should therefore embrace experimentation as a valuable tool for generating empirical evidence, wherever it can be practically and defensibly implemented. This approach should be pursued not as a substitute but as a complement to carefully designed and executed observational studies, with each approach contributing to deepen our understanding of causal processes.

7.3 Quasi-experimental Designs for Analyzing Causal Mediation

In the preceding section, we considered the complexities and challenges of using experimental methods to analyze causal mediation, especially the difficulty of manipulating mediators directly. In social science research, direct manipulation of either the mediator or the exposure is often infeasible. These limitations necessitate exploring quasi-experimental designs that sidestep the challenges of direct manipulation while still addressing concerns about unobserved confounding.

One such approach is the use of randomized encouragement designs, which involve a form of instrumental variable (IV) analysis (Angrist and Krueger 2001; Imbens and Angrist 1994; Imai et al. 2013). In these designs, participants are randomly encouraged or incentivized to adopt specific conditions for the exposure and/or mediator. This strategy allows for indirect manipulation of key variables by way of an instrument–the randomized encouragement. And it enables researchers to relax some of the more stringent assumptions about unobserved confounding that are typical in purely observational studies, while also circumventing the complexities of direct manipulation. Similarly, abrupt or unplanned changes in an exposure or mediator can sometimes create a "natural experiment." By using difference-in-difference (DiD) methods to compare distinct groups of individuals over time—before and after a change in the exposure or mediator—researchers can also relax certain assumptions about unobserved confounding without any direct manipulation of variables (Halaby 2004; Morgan and Winship 2014; Wing et al. 2018).

In this section, we introduce a range of quasi-experimental methods for analyzing causal mediation. We begin by reviewing conventional IV methods aimed at isolating the total effect of an exposure on an outcome, where only the exposure is instrumented in a standard encouragement experiment. Next, we discuss an IV approach where the mediator is instrumented rather than the exposure, as in a parallel encouragement experiment where the exposure is randomly assigned but the mediator is influenced indirectly through a randomized encouragement. Finally, we consider DiD methods for analyzing mediation, initially focusing on standard models for estimating total effects and then introducing models for direct and indirect effects.

Throughout this section, we illustrate these methods by proposing hypothetical modifications to the JOBSII experiment and revisiting our prior analyses of data from this study (Vinokur et al. 1995; Vinokur and Schul 1997; Imai et al. 2010a). Originally, the experiment aimed to determine whether participating in a job training workshop improved subsequent employment by boosting participants' job search self-efficacy—an abstract psychological construct that is not only difficult to measure but also challenging, if not impossible, to manipulate directly. We outline potential modifications to this experiment using randomized encouragement and IV methods, as well as repeated measures and a DiD approach. These adaptations are intended to facilitate inferences about causal mediation under less stringent assumptions about unobserved confounding.

7.3.1 Instrumental Variable (IV) Analyses with an Instrumented Exposure

In a standard encouragement experiment, participants are randomly encouraged—either offered, incentivized, or otherwise motivated—to adopt a specific condition on the exposure. However, the exposure actually received might not align with what was encouraged because participants often fail to comply with or respond to the encouragement as intended. Nevertheless, under certain conditions, this design can identify and consistently estimate an average total effect of the exposure on the outcome among the subpopulation of individuals who comply with the encouragement. Additionally, under more stringent conditions, this design can also recover the average total effect among all individuals in the target population, regardless of their response to the encouragement.

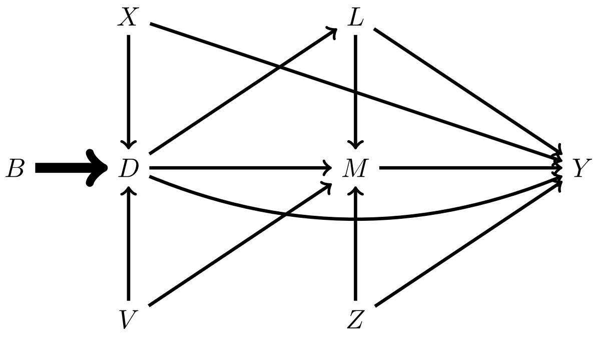


Figure 7.5: A Graphical Mediation Model Depicting Causal Relations in a Standard Encouragement Experiment.

Note: B denotes the randomized encouragement, D denotes the exposure, M denotes a focal mediator, and Y denotes the outcome. X, V, Z and L are unobserved variables that may affect the exposure, mediator, or outcome in different ways.

Figure 7.5 displays a DAG outlining a set of causal relationships in a standard encouragement experiment. In this graph, B denotes the randomized encouragement, coded 1 if a participant is encouraged to adopt the treatment of interest and 0 if not. The variable D captures the exposure actually received, with 1 indicating that a participant was exposed to treatment and 0 otherwise. As before, Y represents the outcome, M represents a focal mediator, and the other variables, denoted by X, V, Z, and L, represent unobserved factors that may confound the the relationships among the exposure, mediator, or outcome. The figure illustrates how random assignment of the encouragement B ensures that it is independent of all potential confounding variables. However, randomizing B alone does not guarantee that the effects of the exposure D are free from confounding, nor does it eliminate any confounding for the effects of the mediator M.

Although the encouragement design does not ensure the absence of unobserved confounding for the exposure, it can still identify and consistently estimate total effects, provided that certain additional conditions are met. Specifically, data from this design can be used with IV methods to identify and estimate an average total effect of the exposure D on the outcome Y among the subpopulation of individuals who comply with the encouragement (Angrist et al. 1996; Imbens and Angrist 1994).

The compliant subpopulation in this context refers to those who take up the treatment when encouraged to do so and who adopt the control condition without such encouragement. Formally, let $D(b)$ denote the potential value of the exposure under encouragement b, where $D(1)$ represents the exposure when a participant is encouraged to take up the treatment and $D(0)$ represents their exposure in the absence of encouragement. With a binary exposure, individuals in the target population can be divided into four distinct subpopulations based on their response to the encouragement:

1. *exposure-compliers*, who adhere to the encouragement by taking up the treatment when encouraged $(D(1) = 1)$ and the control condition when not $(D(0) = 0)$;

2. *exposure-defiers*, who act contrary to the encouragement by opting for the control condition when encouraged to take the treatment $(D(1) = 0)$ and for the treatment when not encouraged $(D(0) = 1)$;

3. *always takers*, who consistently adopt the treatment regardless of the encouragement $(D(1) = 1$ and $D(0) = 1)$; and

4. *never takers*, who consistently opt for the control condition irrespective of the encouragement ($D(1) = 1$ and $D(0) = 1$).

The average total effect of the exposure on the outcome among the subpopulation who comply with the encouragement can then be defined as follows:

$$ATE_{ec}(d, d^*) = \mathbb{E}\left[Y(d) - Y(d^*) \,|D(1) = 1, D(0) = 0\right], \tag{7.11}$$

where the subscript "ec" indicates that this captures an effect specifically for the "exposure-compliers."

This effect can be identified and consistently estimated from a standard encouragement experiment under the following set of assumptions: (g.i) independence of the encouragement with respect to the potential values of the exposure and the outcome, (g.ii) relevance of the encouragement with respect to the exposure, (g.iii) no effect of the encouragement on the outcome except through the exposure, and (g.iv) monotonicity (Angrist et al. 1996; Imbens and Angrist 1994).

Assumption (g.i). The independence assumption requires that the encouragement must be unrelated to the potential values of both the exposure and outcome. This can be expressed formally as follows:

$$\{Y(b, d), D(b)\} \perp B, \tag{7.12}$$

where $\perp$ denotes statistical independence, $D(b)$ is the potential exposure under encouragement b, and $Y(b, d)$ is the joint potential outcome under encouragement b and exposure d. Substantively, this assumption requires that the relationship between the encouragement and both the exposure and outcome must not be confounded by unobserved factors. It would be satisfied with data generated from a process resembling Figure 7.5, where there are no unobserved factors influencing both B and D or B and Y. It is met by design in a standard encouragement experiment, since B is randomly assigned.

Assumption (g.ii). The relevance assumption requires that the encouragement must affect the exposure. Formally, this assumption can be expressed as follows:

$$\mathbb{E}\left[D(b) - D(b^*)\right] \neq 0. \tag{7.13}$$

where b and b^* just denote different values of the encouragement (e.g., 1 and 0 when it is binary). Unlike the independence assumption, the relevance assumption is not met by design in a standard encouragement experiment, but it can be empirically verified. This is accomplished by estimating the effect of the encouragement on the exposure and assessing its size and statistical significance. When this effect is both substantively large and statistically significant at stringent thresholds, the relevance condition is considered satisfied (Bound et al. 1995; Angrist and Pischke 2009).

Assumption (g.iii). Known as the *exclusion restriction*, this assumption requires that the encouragement must not affect the outcome except through its influence on the exposure. It can be formally expressed as follows:

$$Y(b, d) = Y(d), \tag{7.14}$$

which stipulates that the encouragement does not directly affect the outcome nor does it indirectly affect the outcome through any variables other than the exposure. This assumption would be satisfied in data generated from a process resembling Figure 7.5, where the encouragement B affects the outcome Y only through causal paths involving the exposure D. Unlike assumptions (g.i) and (g.ii), however, the exclusion restriction is not met by the design of a standard encouragement experiment and cannot be empirically verified. It would be violated, for example, if the encouragement directly affects a mediator, like M, which in turn influences the outcome. The absence of such paths in Figure 7.5 represents an assumption, not a feature of the encouragement design.

If the assumptions of independence, relevance, and exclusion are all satisfied, then the encouragement B would fulfill the criteria that define a valid *instrumental variable* (IV). Essentially, an IV must be as good as randomly assigned, influence the exposure of interest, and not affect the outcome except through the exposure. In a standard encouragement experiment, the independence condition is met by design, while the relevance condition is testable, but the exclusion restriction remains an unverifiable assumption.

Assumption (g.iv). The last assumption requires that the encouragement affects the exposure monotonically. Formally, this can be expressed as follows:

$$D\left(1\right) \geq D\left(0\right) \text{ for every member of the target population.} \tag{7.15}$$

In substantive terms, the monotonicity assumption stipulates that there must not be any exposure-defiers. Similar to the exclusion restriction, this assumption is neither met by design nor empirically verifiable in a standard encouragement experiment.[1]

If assumptions (g.i) to (g.iv) are all satisfied, along with a standard consistency assumption, the $ATE_{ec}\left(d, d^{*}\right)$ can be identified with the following function of observable data:

$$ATE_{ec}\left(d, d^{*}\right) = \frac{\mathbb{E}\left[Y|B=1\right] - \mathbb{E}\left[Y|B=0\right]}{\mathbb{E}\left[D|B=1\right] - \mathbb{E}\left[D|B=0\right]}. \tag{7.16}$$

The numerator of this expression is the difference in expected values of the outcome between individuals who were encouraged versus not encouraged to adopt the treatment, while the denominator represents the difference in expected values of the exposure between these two groups.

To estimate this quantity, we employ two-stage least squares (TSLS; Angrist et al. 1996; Imbens and Angrist 1994). In the first stage, the exposure D is regressed on the encouragement B using a linear model, which can be expressed as follows:

$$\mathbb{E}\left[D|B\right] = \nu_0 + \nu_1 B. \tag{7.17}$$

After fitting this model by least squares, the second stage of this approach involves regressing the outcome Y on the predicted values for exposure D obtained from Equation 7.17. Specifically, the outcome model in the second stage takes the following form:

$$\mathbb{E}\left[Y|\hat{D}\right] = \omega_0 + \omega_1 \hat{D}, \tag{7.18}$$

where $\hat{D} = \hat{\nu}_0 + \hat{\nu}_1 B$ represents the estimated conditional mean of the exposure D given the encouragement B, as computed from the first stage. Fitting this model by least squares provides an estimate for the total effect among the exposure-compliers, denoted by $\widehat{ATE}_{ec}\left(d, d^{*}\right)^{tsls} = \hat{\omega}_1\left(d - d^{*}\right)$, which is just the estimated coefficient on $\hat{D}$ multiplied by the exposure contrast of interest. The TSLS estimator is consistent provided

[1] There are certain special cases, however, in which the monotonicity assumption can be satisfied by design. For example, in an encouragement experiment where the treatment is a medication that is not yet available commercially, and the encouragement is an offer to receive the drug, those not offered treatment have no alternative means to obtain it, thus eliminating the possibility of any exposure-defiers.

that assumptions (g.i) to (g.iv) are met, and inferential statistics can be constructed using the nonparametric bootstrap. If these assumptions are violated, TSLS is not consistent, and inferences based on the bootstrap are invalid.

Under an additional assumption, the same IV approach can be used to recover the average total effect among the entire target population, not merely among the subpopulation of exposure-compliers. Specifically, the average total effect can be identified and consistently estimated under assumptions (g.i) to (g.iii), along with an additional assumption (g.v) that the effects of the exposure on the outcome are homogeneous across subpopulations defined in terms of their response to the encouragement. By substituting the monotonicity assumption with this stronger assumption about effect homogeneity, we can extend our analysis from the $ATE_{ec}(d, d^*)$ to the $ATE(d, d^*)$.

Assumption (g.v). Formally, the homogeneity assumption can be expressed as follows:

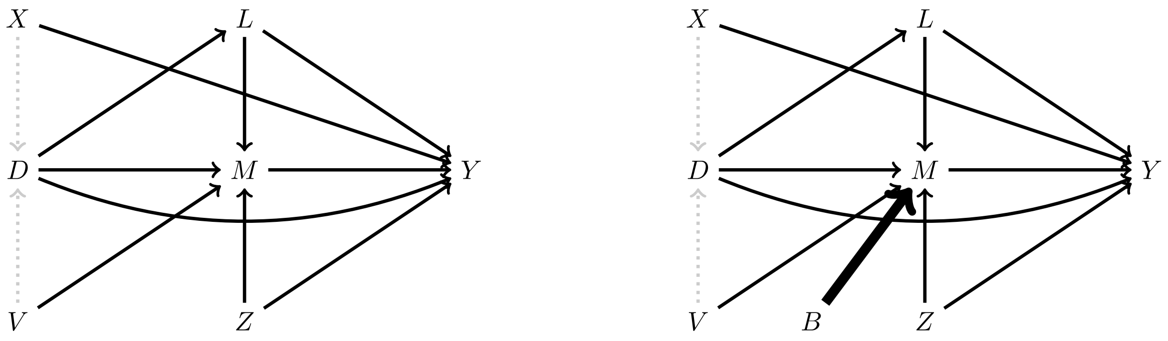
$$\mathbb{E}\left[Y(d) - Y(d^*)\,|D(b), D(b^*)\right] = \mathbb{E}\left[Y(d) - Y(d^*)\right] \tag{7.19}$$

It requires that the total effects of the exposure on the outcome are invariant across the exposure-compliers, exposure-defiers, always takers, and never takers. This is a strong assumption that is not met by design in a standard encouragement experiment and cannot be empirically verified. Nevertheless, if it holds, along with the assumptions of independence, relevance, and exclusion, then the $ATE(d, d^*)$ can be identified and consistently estimated using the same methods outlined previously for the $ATE_{ec}(d, d^*)$ because, under homogeneity, these effects are equivalent.

To summarize, with data from a standard encouragement experiment, the analysis of total effects is no longer predicated on unverifiable assumptions about the absence of unobserved confounding between the exposure and outcome, as outlined in Chapter 3. Instead, it necessitates several alternative assumptions that also cannot be verified: namely, that the encouragement only affects the outcome through its influence on the exposure and that this influence on the exposure is monotonic. These conditions are essential for drawing valid inferences about the total effect among the subpopulation of exposure-compliers. To extend these inferences to the entire target population, the monotonicity assumption must be supplanted with an even stronger assumption that the average effects of the exposure on the outcome are invariant across subpopulations who respond differently to the encouragement.

We focused on the application of IV methods in the context of standard encouragement experiments, where the exposure is indirectly manipulated through an instrument that is randomly assigned by the researcher. This design ensures that the independence assumption is satisfied, which then allows for empirical confirmation of the relevance assumption as well. IV methods can also used in observational studies, where the instruments are purported to arise naturally. However, in these settings, there is no guarantee that the independence assumption is satisfied, nor can the relevance assumption be rigorously tested. Consequently, when applying IV methods with observational data, the assumptions of independence and relevance must also be treated as unverifiable. Along with the exclusion restriction, monotonicity, and homogeneity, then, these assumptions require the same level of critical scrutiny that is often reserved for concerns about unobserved confounding.

In general, IV methods are most compelling when the instrument is directly assigned at random by the researcher. They can also generate credible inferences when the instrument arises from a natural experiment, where variation in the instrument is due to a stochastic process known but not necessarily controlled by the researcher (e.g., a school enrollment or military draft lottery). In IV analyses of purely observational data, credibility can sometimes be enhanced by adjusting for a set of covariates C. When adjusting for



A. Arm One of the Experiment

B. Arm Two of the Experiment

Figure 7.6: A Graphical Mediation Model Depicting Causal Relations in a Parallel Encouragement Experiment.

Note: B denotes the randomized encouragement, D denotes the exposure, M denotes a focal mediator, and Y denotes the outcome. X, V, Z and L are unobserved variables that may affect the exposure, mediator, or outcome in different ways. The dotted, faded arrows denote causal paths that are broken via direct randomization of the exposure.

covariates, the independence and relevance assumptions must now hold conditionally on these variables, such that $\{Y(b,d), D(b)\} \perp B|C$ and $\mathbb{E}[D(b) - D(b^*)|C] \neq 0$. To estimate the total effects of interest, these covariates are then included as additional predictors in both the first and second stage of the TSLS approach. Aside from this modification, the analysis proceeds exactly as outlined previously (Angrist and Pischke 2009, 2014).

7.3.2 IV Analyses with an Instrumented Mediator

In a parallel encouragement design, participants are randomly assigned to one of two experimental arms. In the first arm, the exposure alone is directly randomized without manipulating the mediator, similar to a standard experiment. In the second arm, the exposure is also directly randomized, but the mediator is manipulated indirectly through a randomized encouragement. The encouragement is intended to induce participants to adopt higher versus lower values on the mediator. The first arm of the experiment satisfies all the assumptions necessary for nonparametrically identifying and consistently estimating the average total effect, as outlined in Section 7.2.2. Moreover, the second arm can identify and consistently estimate the natural direct effect under certain conditions, and by combining data from both arms of the encouragement experiment, the natural indirect effect can be recovered as well.

Figure 7.6 presents a DAG depicting a set of causal relationships in a parallel encouragement experiment. Panel A shows the causal relationships in the first arm of the experiment, where only the exposure D is directly randomized, thereby eliminating any potential confounding of its effects on the outcome Y. Panel B depicts causal relationships in the second arm, where the exposure D is again directly randomized, while the mediator M is manipulated indirectly through a randomized encouragement. This encouragement, denoted by the variable B, is coded 1 if a participant is encouraged to adopt a higher value of the mediator, and 0 otherwise. The other variables in the figure, denoted by X, V, Z, and L, represent unobserved factors. The graph illustrates that random assignment of the exposure D and encouragement B ensures their independence from all potentially confounding variables, but it does not guarantee that the effects of the mediator M itself are unconfounded.

As in standard experiments and parallel randomization designs, random assignment of the exposure in the first arm of a parallel encouragement design satisfies all the conditions necessary for identifying the average total effect, $ATE\left(d, d^{*}\right)=\mathbb{E}\left[Y(d)-Y\left(d^{*}\right)\right]$. Moreover, this effect can be consistently estimated simply by comparing sample means of the outcome among participants assigned to different levels of the exposure. Specifically, if $A=1$ denotes assignment to the first arm of the experiment, $A=2$ denotes assignment to the second arm, and $\bar{Y}_{d,a}$ represents the sample mean of the outcome for participants assigned to exposure $D=d$ in arm $A=a$, then the average total effect can be estimated as $\widehat{ATE}\left(d, d^{*}\right)^{pex}=\bar{Y}_{d,1}-\bar{Y}_{d^{*},1}$. The superscript “pex” indicates that this estimator is based on a parallel encouragement experiment. However, it is equivalent to the corresponding estimator from a parallel randomization experiment, as described in Section 7.2.2, since the first arms of these two designs are identical.

In the second arm, random assignment of the exposure guarantees the absence of exposure-mediator and exposure-outcome confounding, but the randomized encouragement does not eliminate the possibility of exposure-induced or unobserved confounding for the mediator-outcome relationship. Nevertheless, data from this arm can be used with IV methods to identify and consistently estimate the natural direct effect of the exposure on the outcome, $NDE\left(d, d^{*}\right)=\mathbb{E}\left[Y\left(d, M\left(d^{*}\right)\right)-Y\left(d^{*}, M\left(d^{*}\right)\right)\right]$, provided that certain conditions are met. Specifically, this effect can be identified and consistently estimated under the following assumptions: (h.i) independence of both the exposure and the encouragement with respect to the potential values of the mediator and the outcome, (h.ii) relevance of the encouragement with respect to the mediator, (h.iii) no effect of the encouragement on the outcome except through the mediator, (h.iv) homogeneity of the joint effects of the exposure and mediator, and (h.v) no exposure-mediator interaction effects. These assumptions are similar to those outlined previously for an IV analysis of a standard encouragement experiment. However, in this context, the assumptions of independence, relevance, and exclusion now apply to the mediator, and identifying natural effects requires restrictions on both subpopulation heterogeneity and effect interaction.

Assumption (h.i). The independence assumption here requires that both the exposure and the encouragement are unrelated to the potential values of the mediator and outcome in the second arm of the experiment. Formally, this assumption can be expressed as follows:

$$\{Y(b, d, m), M(b, d)\} \perp\{B, D\}, \tag{7.20}$$

where $M(b, d)$ is the potential value of the mediator under encouragement b and exposure d, and $Y(b, d, m)$ is the potential outcome under encouragement b, exposure d, and a level of the mediator given by m. Substantively, this assumption requires that the relationships between the exposure and the mediator, as well as the exposure and the outcome, are not confounded by unobserved factors. In addition, it requires that the relationship of the encouragement with both the mediator and the outcome must also be unconfounded. This assumption is met by design in the second arm of a parallel encouragement experiment, where both the exposure D and encouragement B are randomly assigned, as depicted in Figure 7.6.

Assumption (h.ii). Under a parallel encouragement design, the relevance assumption requires that the encouragement must affect the mediator, which can be formally expressed as follows:

$$\mathbb{E}\left[M(b, d)-M\left(b^{*}, d\right)\right] \neq 0. \tag{7.21}$$

The relevance of the encouragement to the mediator is not guaranteed in a parallel encouragement experiment. Instead, it must be empirically verified by assessing both the practical and statistical significance of the relationship between the encouragement and mediator in the second arm of the experiment. This assessment should mirror the procedures used to evaluate similar assumptions in standard IV analyses (Angrist and Pischke 2009, 2014).

Assumption (h.iii). The exclusion restriction in a parallel encouragement design requires that the encouragement should not affect the outcome except through its influence on the mediator, specifically in the second arm of the experiment. Formally, this assumption can be expressed as follows:

$$Y\left(b, d, m\right) = Y\left(d, m\right), \tag{7.22}$$

which implies that the encouragement does not directly or indirectly affect the outcome through any variables other than the mediator of interest. This assumption is satisfied in Figure 7.6, where the encouragement B influences the outcome Y solely through the causal path $B \rightarrow M \rightarrow Y$. However, the exclusion restriction is not met by design in a parallel encouragement experiment and cannot be empirically verified. It would be violated, for example, if the encouragement directly influenced another mediator besides M. This is a serious concern in parallel encouragement experiments, especially those in which the focal mediator is difficult to manipulate, thereby necessitating a strong or intrusive encouragement that might impact other related mediators. Nevertheless, if assumptions (h.i) to (h.iii) are all met, then the encouragement B is a valid instrument for the mediator M.

Assumption (h.iv). Identifying the natural direct effect in a parallel encouragement design also requires an additional assumption that the joint effects of the exposure and mediator on the outcome are invariant across subpopulations defined by how their mediator responds to the encouragement. This version of the homogeneity assumption can be formally expressed as follows:

$$\mathbb{E}\left[Y\left(d, m\right) - Y\left(d^*, m^*\right) | M\left(b\right), M\left(b^*\right)\right] = \mathbb{E}\left[Y\left(d, m\right) - Y\left(d^*, m^*\right)\right]. \tag{7.23}$$

It is not guaranteed by the design of a parallel encouragement experiment and cannot be empirically evaluated.

Assumption (h.v). The final assumption necessary for identifying the natural direct effect in a parallel encouragement design stipulates that the effects of the exposure and mediator do not interact to influence the outcome. Formally, this assumption can be expressed as follows:

$$Y\left(d, m\right) - Y\left(d^*, m\right) = Y\left(d, m^*\right) - Y\left(d^*, m^*\right) \text{ for every member of the target population.} \tag{7.24}$$

As discussed in Section 7.2.2, the assumption that there is no interaction between the exposure and mediator must hold for every individual, not merely on average for the target population as a whole. Consequently, it also cannot be empirically confirmed.

If assumptions (h.i) to (h.v) are all met, along with a standard consistency assumption, the $NDE\left(d, d^*\right)$ can be consistently estimated using TSLS with data from the second arm of a parallel encouragement design. To implement this approach, the mediator M is first regressed on the encouragement B and exposure D in a linear model with the following form:

$$\mathbb{E}\left[M | B, D, A = 2\right] = \alpha_0 + \alpha_1 B + \alpha_2 D. \tag{7.25}$$

Then, the outcome Y is regressed on the exposure D and the predicted values for the mediator M, which are obtained from least squares estimates of Equation 7.25. Specifically, the outcome model can be expressed as follows:

$$\mathbb{E}\left[Y|D,\hat{M},A=2\right]=\beta_0+\beta_1 D+\beta_2\hat{M}, \tag{7.26}$$

where $\hat{M}=\hat{\alpha}_0+\hat{\alpha}_1 B+\hat{\alpha}_2 D$ represents the estimated conditional mean of the mediator M given the encouragement B and exposure D. Fitting the outcome model by least squares provides an estimate for the natural direct effect, denoted as $\widehat{NDE}\left(d,d^*\right)^{pex}=\hat{\beta}_1\left(d-d^*\right)$, which is equal to the estimated coefficient on D multiplied by the exposure contrast of interest.

Furthermore, by obtaining estimates of the total and natural direct effects from the first and second arms of the experiment, respectively, we can derive an estimate for the natural indirect effect. Specifically, an estimator for the natural indirect effect is given by $\widehat{NIE}\left(d,d^*\right)^{pex}=\widehat{ATE}\left(d,d^*\right)^{pex}-\widehat{NDE}\left(d,d^*\right)^{pex}$, which represents the difference between the total effect estimated in the first arm and the TSLS estimate for the natural direct effect in the second arm. The consistency of these estimators, $\widehat{NDE}\left(d,d^*\right)^{pex}$ and $\widehat{NIE}\left(d,d^*\right)^{pex}$, hinges on the assumptions that B is a valid instrument for M and that the effects of the exposure and mediator are homogeneous and do not interact.

When the homogeneity assumption is met but the assumption of no exposure-mediator interaction is violated, it is still possible to identify and consistently estimate controlled direct and eliminated effects, though not natural effects. In this case, the TSLS approach can be modified to include an interaction term between the exposure and encouragement in the first stage, which serves as an additional instrument for the mediator. This model would then be specified as follows:

$$\mathbb{E}\left[M|B,D,A=2\right]=\alpha_0+\alpha_1 B+D\left(\alpha_2+\alpha_3 B\right). \tag{7.27}$$

Next, in the second stage, the outcome Y is regressed on the exposure D, the predicted values for the mediator M from the first stage, and an interaction between these terms. This model can now be expressed as follows:

$$\mathbb{E}\left[Y|D,\hat{M},A=2\right]=\beta_0+\beta_1 D+\hat{M}\left(\beta_2+\beta_3 D\right), \tag{7.28}$$

where $\hat{M}=\hat{\alpha}_0+\hat{\alpha}_1 B+D\left(\hat{\alpha}_2+\hat{\alpha}_3 B\right)$. Fitting these model successively yields a TSLS estimate for the controlled direct effect, denoted as $\widehat{CDE}\left(d,d^*,m\right)^{pex}=\hat{\beta}_1+\hat{\beta}_3 m$. By extension, an estimate from the eliminated effect can be obtained by computing $\widehat{AEE}\left(d,d^*,m\right)^{pex}=\widehat{ATE}\left(d,d^*\right)^{pex}-\widehat{CDE}\left(d,d^*,m\right)^{pex}$, the difference between the total effect estimated in the first arm and the TSLS estimate for the controlled direct effect in the second arm.

When the homogeneity assumption is violated as well, it may still be possible to analyze a variant of the controlled direct effect among the subpopulation of mediator-compliers, that is, among individuals who adopt a higher value of the mediator when encouraged and a lower value when not encouraged. Controlled direct effects among mediator-compliers can be identified and estimated even when homogeneity does not hold, provided that the encouragement affects the mediator monotonically, such that $M\left(1,d\right)\geq M\left(0,d\right)$ for all individuals. However, the difference between this variant of the controlled direct effect and the total effect estimated from the first arm does not provide a meaningful estimate for the eliminated effect. This discrepancy arises because the total effect and controlled direct effect now pertain to different populations. The total effect represents the entire target population as a whole, while the controlled direct effect pertains only to the subpopulation of mediator-compliers. These effects may differ because controlling the mediator at a fixed value could either attenuate or amplify the exposure's effect on the outcome. Alternatively, they may differ simply because the effects of the exposure or mediator vary across subpopulations defined by their response to the encouragement. In general, relaxing the homogeneity assumption in IV analyses of causal mediation introduces additional complexities that are challenging to resolve (Bullock and Ha 2011).

To illustrate these methods, let's revisit the JOBSII study (Vinokur et al. 1995; Vinokur and Schul 1997; Imai et al. 2010a). This study involved a standard experiment designed to evaluate the impact of a job training workshop on subsequent employment. Participants were randomly assigned either to the workshop ($D = 1$) or a control group ($D = 0$). The mediator of interest–job search self-efficacy, denoted by M–was merely observed, not manipulated. And the outcome, denoted by Y, captures whether a participant secured employment after the workshop.

In Chapter 3, we performed a mediation analysis on these data. The validity of this analysis depended on strong assumptions about the absence of unobserved and exposure-induced confounding for the relationship between the mediator and outcome. These assumptions are not met by the design of the original experiment. Moreover, they are subject to a nontrivial measure of doubt, since there are many unobserved factors that may influence both self-efficacy and employment, such as a participant's disability status or personality.

Suppose, however, that the JOBSII study were adapted to include another experimental arm. In this second arm, participants would also be randomly assigned to either the workshop or a control group. Additionally, they would receive a randomized encouragement, denoted by B, designed to indirectly manipulate the mediator, job search self-efficacy. Self-efficacy is a complex psychological construct, but it could conceivably be influenced by an autobiographical memory task (AMT; Imai et al. 2013; Mills and D'Mello 2014; Rubin 2005), a method often used to manipulate emotions, feelings, and other psychological states. In our hypothetical redesign of the JOBSII study, participants in the second arm would be randomly assigned to complete the AMT after the workshop. Those receiving the AMT ($B = 1$) would be asked to recall and write about a past experience that boosted their confidence in their ability to find work. Those not assigned the AMT ($B = 0$) would not engage in this recollection and writing task. The researchers would then follow up with participants later on to measure their job search self-efficacy and subsequent employment.

Under this experimental design, the average total effect of the workshop on employment can be identified and consistently estimated in the first arm, following the approach used in the original study. Specifically, the average total effect of the workshop on employment is estimated by comparing the employment rate for participants assigned to the workshop in the first arm of the experiment and the employment rate for those in the control group of the first arm.

In the second arm of this experiment, the natural direct effect of the workshop on employment could be identified and consistently estimated if the AMT is a valid instrument for job search self-efficacy and if the joint effects of the workshop and self-efficacy are homogeneous and do not interact. The AMT would serve as a valid instrument if it satisfies the assumptions of independence (h.i), relevance (h.ii), and exclusion (h.iii). The independence assumption is guaranteed by design, as the encouragement was randomly assigned. The relevance assumption can be empirically verified by examining the association of the AMT with job search self-efficacy. And the exclusion restriction would hold as along as the AMT does not affect employment, except through its impact on job search self-efficacy. This is a strong, unverifiable assumption that could be violated if completing the AMT influences other psychological states that shape the prospects of securing future employment. For example, if the AMT not only boosts self-efficacy but also reduces anxiety, and less anxious participants are more successful in job interviews, then the exclusion restriction would not be satisfied. The assumptions of homogeneity and no exposure-mediator interaction are also not guaranteed by design and cannot be empirically verified. They would be violated if the effects of the workshop and self-efficacy interact to influence employment for any individual in the target population, or if the joint effects of these variables differ across subpopulations defined by their response to the AMT.

By adapting the JOBSII experiment to follow a parallel encouragement design, we can now estimate the natural direct effect in the second arm using the AMT as an instrument for self-efficacy. This approach relies on assumptions about the validity of the instrument and the absence of heterogeneous or interactive effects, rather than assumptions about the absence of unobserved or exposure-induced confounding. If these alternative assumptions are satisfied, the natural direct effect could be estimated by TSLS in the second arm of the experiment. The first stage would involve regressing job search self-efficacy on indicators of assignment to the workshop and the AMT. After this model is fit by least squares, the second stage would involve regressing employment status on the indicator of workshop participation and the predicted values for self-efficacy from the first stage. A TSLS estimate for the natural direct effect is given by the coefficient on the workshop indicator in this outcome regression. In addition, an estimate for the natural indirect effect could then be obtained by computing the difference between the estimated total effect from the first arm and the TSLS estimate of the direct effect from the second arm. Confidence intervals and hypothesis tests could be constructed using the nonparametric bootstrap.

Although using a parallel encouragement design for mediation analysis with an instrumental variable obviates the need for assumptions about unobserved or exposure-induced confounding, several limitations persist. In particular, IV methods still depend on strong, unverifiable assumptions, just not ones that concern unobserved confounding. For example, while randomly assigning the encouragement satisfies the independence and relevance assumptions, the exclusion restriction requires that the encouragement affects only the targeted mediator and no others. This requirement poses significant challenges in practice, as it necessitates interventions that selectively influence a single mediator without affecting other potential mediators. Achieving such precision is often difficult and unrealistic in social science applications, especially when the mediators involve complex psychological phenomena.

Another significant limitation arises when the effects of interest are heterogeneous or interact. Even a well-designed experiment may yield inaccurate inferences about causal mediation if the joint effects of the exposure and mediator vary among individuals or interact to influence the outcome. The assumption of no exposure-mediator interaction is particularly strong and often untenable, leading to potential biases when estimating natural direct and indirect effects from a parallel encouragement design. While focusing more narrowly on effects within certain latent subpopulations can mitigate some of these concerns, this approach limits the generalizability of findings and raises questions about the theoretical relevance of the alternative estimands. Is it really advisable to prioritize effects among exposure-compliers or mediator-compliers, rather than attempt to understand the causal process in the target population as a whole? It is seldom evident that substituting a more convenient but less relevant estimand for a less convenient but more pertinent one is a prudent analytic strategy.

In summary, parallel encouragement designs analyzed with IV methods present a potentially more robust alternative for empirically evaluating causal mediation. This approach effectively addresses concerns about unobserved and exposure-induced confounding but introduces its own set of contentious assumptions and practical challenges. Researchers must critically assess these limitations and make every effort to carefully design studies that minimize their impact, ensuring that the inferences drawn are both valid and relevant.

7.3.3 Difference-in-difference (DiD) Models for Total Effects

In a pre/post design, measures of the outcome are taken before and after participants are potentially exposed to a treatment of interest. These designs are applicable to both experimental and non-experimental studies. In an experimental pre/post design, a "pre-test" is first administered to measure the outcome at baseline. Subsequently, participants are randomly assigned to either the treatment or control condition, and then a "post-test" is used to measure the outcome following treatment assignment. The total effect of the exposure on the outcome can be consistently estimated using a DiD approach, which involves comparing the difference in outcomes between the pre-test and post-test across the treatment and control groups (Bonate 2000).

In contrast, a non-experimental pre/post design does not involve random assignment of the exposure. Instead, researchers initially collect measurements of the outcome before any exposure to the treatment of interest. Participants then experience the treatment or control condition naturally, without researcher intervention, which is followed by another measurement of the outcome. Although the exposure is not randomized, the total effect can still be identified and consistently estimated using DiD methods, provided that any unobserved confounding of the exposure-outcome relationship is stable over time (Card and Krueger 1994; Halaby 2004; Meyer 1995).

Let Y_t denote the measurement of the outcome taken at time $t \in \{0, 1\}$, where Y_0 is the baseline measurement taken before exposure to the treatment and Y_1 is the follow-up measurement collected after exposure. Similarly, let D_t represent a binary indicator of exposure, coded 1 if a participant is exposed to the treatment at time t and 0 otherwise. By definition, $D_0 = 0$ for all participants, because no one is exposed to treatment at baseline. At follow-up, $D_1 = 0$ indicates exposure to the control condition and $D_1 = 1$ indicates exposure to treatment between the two time points. Using this notation, the average total effect of the exposure on the outcome can be expressed as $ATE\left(d, d^*\right) = \mathbb{E}\left[Y_1\left(d\right) - Y_1\left(d^*\right)\right]$, where $Y_1\left(d\right)$ is the potential outcome at time $t = 1$ under exposure d during the same period.

This effect can be identified and consistently estimated from a pre/post design under the assumptions of consistency and parallel trends. The consistency assumption requires that $Y_t = Y_t\left(D_t\right)$, which implies that the observed outcome at each time point must correspond with a participant's potential outcome under their observed exposure at the same time period. The parallel trends assumption can be formally expressed as $\Delta Y\left(d\right) \perp \Delta D$, where $\Delta D = D_1 - D_0$ represents the change in exposure over time, and $\Delta Y\left(d\right) = Y_1\left(d\right) - Y_0\left(0\right)$ denotes the change in potential outcomes between the two time points. This assumption requires that the change in potential outcomes must be statistically independent of changes in the exposure. Substantively, it implies that if the treatment had not been administered to the treated group, their trend in outcomes would mirror that observed in the control group. Conversely, it also implies that if the control group had received the treatment, their trend in outcomes would resemble that observed in the treated group. This assumption is guaranteed to hold in an experimental pre/post design where the exposure is randomly assigned, as randomization ensures that the change in exposure status is independent of the potential outcomes at each time point and any differences therein.

By contrast, in studies where the exposure is not randomized and participants self-select into treatment, the parallel trends assumption is neither guaranteed nor empirically verifiable. This assumption is untestable because we cannot evaluate trends in the counterfactual outcomes, only the observed outcomes. However, researchers can gauge the validity of this assumption by examining pre-treatment trends in the observed outcomes, if multiple pre-treatment measures are available (Angrist and Pischke 2009; Morgan and Winship 2014). When the treatment and control groups exhibit similar trends in the outcome across multiple pre-treatment measures, it lends some credibility to the parallel trends assumption, even though it may still fail to hold with respect to the counterfactual post-treatment outcomes.

The parallel trends assumption would be violated, for example, if there are any time-varying confounders–factors that influence both the outcome and selection into treatment, and that vary from one period to the next. This assumption would also be compromised by time-invariant confounders whose effects on the outcome differ over time. These are factors that influence selection into treatment and impact the outcome differently across time periods, even if the variables themselves remain unchanged. Finally, the parallel trends assumption could also be violated if there are carry-over effects of the baseline outcome on either the exposure or the follow-up measurement of the outcome. In other words, if the baseline outcome influences selection into treatment and/or directly affects the outcome measured at follow-up, parallel trends may not be satisfied.

Nevertheless, the parallel trends assumption does not preclude *all* forms of confounding. Specifically, if there are confounding factors that contaminate the exposure-outcome relationship, but they are time-invariant and their effects on the outcome also remain stable over time, the parallel trends assumption will still hold. This implies that average total effects can be identified under relatively weaker assumptions about confounding in non-experimental pre/post designs, compared to typical observational studies with a cross-sectional design. In a typical observational study, all forms of unobserved confounding must be absent. However, in a pre/post design, identifying the total effect requires only the absence of time-varying confounders, time-invariant confounders whose effects on the outcome differ across time, and carry-over effects of the baseline outcome. Thus, if the only confounding influences are temporally stable, total effects can still be identified in a non-experimental pre/post design, even if these factors are unobserved. Essentially, a pre/post design allows for the identification of total effects in the presence of time-invariant confounding, whether observed or not.

If the parallel trends and consistency assumptions are satisfied in a pre/post design, the $ATE\left(d, d^{*}\right)$ can be identified with the following function of observable data:

$$ATE\left(d, d^{*}\right)=\mathbb{E}\left[\Delta Y \mid \Delta D=d\right]-\mathbb{E}\left[\Delta Y \mid \Delta D=d^{*}\right], \tag{7.29}$$

where $\Delta Y = Y_1 - Y_0$ is the difference in outcomes between the baseline and follow-up measurements, and ΔD is defined analogously (Angrist and Pischke 2009; Morgan and Winship 2014). The DiD approach derives its name from this expression. Specifically, the first term in Equation 7.29 is the expected value of the pre/post difference in outcomes among those exposed to d, while the second term represents the expected difference in outcomes over time among those exposed to d^{*}. The total effect is identified by the difference between these quantities, hence the label "difference in differences."

To clarify how the DiD approach resolves the problem of time-invariant confounding, the expression in Equation 7.29 can be rewritten as follows:

$$ATE\left(d, d^{*}\right)=\left(\mathbb{E}\left[Y_{1} \mid \Delta D=d\right]-\mathbb{E}\left[Y_{1} \mid \Delta D=d^{*}\right]\right)-\left(\mathbb{E}\left[Y_{0} \mid \Delta D=d\right]-\mathbb{E}\left[Y_{0} \mid \Delta D=d^{*}\right]\right). \tag{7.30}$$

In this expression, the first term represents the expected difference in outcomes at follow-up between the treatment and control groups, while the second term captures the expected difference in baseline outcomes comparing these groups. The difference in baseline outcomes serves as an indicator of confounding bias. Because the outcome measurement at baseline occurs before treatment is administered, any differences observed between the treatment and control groups at this point in time are attributable to confounding factors other than the treatment itself. If the bias due to these confounding factors remains constant from one period to the next, then subtracting the difference in baseline outcomes from the difference in follow-up outcomes isolates the effect of treatment (Hernan and Robins 2020).

To estimate the total effect, the difference in outcomes can be regressed on the difference in exposures using a linear model with the following form:

$$\mathbb{E}\left[\Delta Y | \Delta D\right] = \gamma_0 + \gamma_1 \Delta D. \tag{7.31}$$

Fitting this model by least squares provides an estimate for the total effect, which is given by $\widehat{ATE}\left(d, d^*\right)^{DiD} = \hat{\gamma}_1\left(d - d^*\right)$, the estimated coefficient on ΔD multiplied by the exposure contrast of interest. The "DiD" superscript denotes that this is a "difference in differences" estimator. This estimator will be consistent for the average total effect provided that the parallel trends assumption holds–that is, as long as there are no unobserved confounders that vary over time or whose effects on the outcome differ from one period to the next. Inferential statistics can be generated using the nonparametric bootstrap, resampling participants with replacement for each replication.

To summarize, the analysis of total effects in a pre/post design does not rely on strong, unverifiable assumptions about the complete absence of unobserved confounding between the exposure and outcome. Instead, it merely requires the absence of unobserved confounding by factors that either vary over time or are time-invariant but whose influence on the outcome changes over time. These conditions are met in an experimental pre/post design, where the exposure is randomized. However, in non-experimental settings without random assignment, they still represent strong, unverifiable assumptions, but they are arguably less stringent compared to those required in observational studies without a pre/post design. Essentially, the pre/post design allows us to relax the assumptions necessary for identifying total effects by accommodating the presence of time-invariant confounding.

The DiD approach can be adapted for more complex study designs and data structures, enhancing its utility and flexibility. For example, it is possible to incorporate controls for observed time-varying confounders, which can mitigate potential biases and bolster the credibility of the parallel trends assumption (Halaby 2004). DiD analyses can also be extended beyond two time points to include multiple pre-treatment and post-treatment measures, and they can be modified to accommodate staggered exposures, where participants receive the treatment at different times (Callaway and Sant'Anna 2021). While these extensions introduce additional complexities that pose other challenges for drawing valid inferences (de Chaisemartin and D'Haultfoeuille 2020; Goodman-Bacon 2021; Imai and Kim 2021), they significantly expand the applicability of the DiD approach, making it a more versatile tool for analyzing causal effects in non-experimental pre/post designs.

7.3.4 DiD Models for Direct and Indirect Effects

Now consider an experimental pre/post design where measurements of both a focal mediator and the outcome are recorded before and after a potential exposure to treatment. Initially, a "pre-test" is used to collect data on the mediator and outcome at baseline, and participants are then randomly assigned to the treatment or control group, with a "post-test" measuring the same variables again following random assignment. In a non-experimental variant of this design, the only difference is that participants self-select into the treatment or control condition between the baseline and follow-up assessments, without any intervention from the researchers. By using a DiD approach with this study design, we can identify and consistently estimate direct and indirect effects under alternative, and arguably less restrictive, assumptions about unobserved confounding, compared to a typical study that lacks repeated measurements.

Let M_t and Y_t represent measurements of the mediator and outcome, respectively, taken at time t, where $t = 0$ denotes the baseline and $t = 1$ denotes the follow-up measurement. As in Section 7.3.3, D_t is a binary indicator of exposure to treatment. At baseline, D_0 is coded 0 for all participants because nobody has yet been exposed to treatment, whereas at follow-up, D_1 is coded 1 if a participant was exposed to treatment in the interim, and 0 otherwise. Using this notation, the natural direct effect of the exposure on the outcome can be expressed as $NDE\left(d, d^*\right) = \mathbb{E}\left[Y_1\left(d, M_1\left(d^*\right)\right) - Y_1\left(d^*, M_1\left(d^*\right)\right)\right]$, the natural indirect effect is given by $NIE\left(d, d^*\right) = \mathbb{E}\left[Y_1\left(d, M_1\left(d\right)\right) - Y_1\left(d, M_1\left(d^*\right)\right)\right]$, and the controlled direct effect is defined as $CDE\left(d, d^*, m\right) = \mathbb{E}\left[Y_1\left(d, m\right) - Y_1\left(d^*, m\right)\right]$. In these expressions, $Y_1\left(d, M_1\left(d\right)\right)$ is the nested potential outcome under exposure d at time $t = 1$, with $Y_1\left(d^*, M_1\left(d^*\right)\right)$ and $Y_1\left(d, M_1\left(d^*\right)\right)$ defined analogously. Similarly, $Y_1\left(d, m\right)$ is the joint potential outcome under level d of the exposure and level m of the mediator at time $t = 1$.

A pre/post design is capable of identifying and consistently estimating these effects, provided that certain conditions are met. Specifically, the $NDE\left(d, d^*\right)$, $NIE\left(d, d^*\right)$, and $CDE\left(d, d^*, m\right)$ can be identified and consistently estimated with data from a pre/post design under the following set of assumptions: (j.i) parallel trends in the outcome with respect to the exposure, (j.ii) parallel trends in the mediator with respect to the exposure, (j.iii) parallel trends in the outcome with respect to the mediator, and (j.iv) invariant effects of the mediator on the outcome.

Assumption (j.i). The assumption of parallel trends in the outcome with respect to the exposure can be formally expressed as follows:

$$\Delta Y\left(d, m\right) \perp \Delta D, \tag{7.32}$$

where $\Delta D = D_1 - D_0$ denotes the change in exposure over time and $\Delta Y\left(d, m\right) = Y_1\left(d, m\right) - Y_0\left(0, M_0\right)$ denotes the change in the joint potential outcomes. This assumption states that the change in potential outcomes must be statistically independent of the changes observed in the exposure. It implies that, had the treatment not actually been administered, the outcome trend in the group assigned to receive the treatment would mirror that observed in the control group. It also implies that, had the control group actually received the treatment, their outcome trends would follow those observed in the treated group.

Assumption (j.ii). The assumption of parallel trends in the mediator with respect to the exposure can be similarly expressed as follows:

$$\Delta M\left(d\right) \perp \Delta D, \tag{7.33}$$

where $\Delta M\left(d\right) = M_1\left(d\right) - M_0\left(0\right)$ denotes the change in the potential values of the mediator over time. This assumption states that the change in potential values of the mediator must be statistically independent of changes in the exposure. Substantively, it implies that, absent the treatment, the trend in the mediator among treated participants would mirror the trend observed in the control group. It also implies that if the control group had actually received the treatment, their trend in the mediator would resemble that observed in the treated group.

In an experimental pre/post design where the exposure is randomly assigned, assumptions (j.i) and (j.ii) are guaranteed to hold. Randomization ensures that the exposure is independent of the potential values of the mediator and the outcome, as well as any changes in these variables over time. However, these assumptions need not hold in a non-experimental pre/post design, where participants self-select into the treatment and control groups between the baseline and follow-up assessments.

Assumption (j.iii). The assumption of parallel trends in the outcome with respect to the mediator can be formally expressed as follows:

$$\Delta Y\left(d,m\right) \perp \Delta M|\Delta D, \tag{7.34}$$

where $\Delta M = M_1 - M_0$ represents the change in the observed values of the mediator over time. It states that the change in the joint potential outcomes must be statistically independent of changes in the mediator, conditional on the change in exposure. This variant of the parallel trends assumption is not guaranteed by the design of an experimental pre/post study, in which the exposure is randomized but not the mediator. By extension, it also need not hold in a non-experimental pre/post design, where neither variable is randomized.

Figure 7.7 displays a series of DAGs outlining different sets of causal relationships that may arise in a pre/post design. In Panel A, the effects of the exposure D_1 and mediator M_1 on the outcome Y_1 at follow-up are confounded only by a time-invariant characteristic, denoted by U. If the effects of U on the mediator and outcome are the same at both times $t = 0$ and $t = 1$, then the parallel trends assumptions discussed previously would all be satisfied in data generated from a process resembling this panel.

However, the other panels of Figure 7.7 illustrate scenarios where these assumptions might not hold. For example, Panel B introduces a time-varying characteristic, denoted by X_t, that confounds the effects of the exposure and mediator on the outcome. In Panel C, the mediator and outcome at baseline have carry-over effects on the exposure, mediator, and outcome measured at follow-up. Panel D incorporates an exposure-induced confounder, denoted by L_t, for the effect of the mediator on the outcome. In this scenario, the change in exposure between time periods induces a change in another intermediate variable, which then influences both the outcome and mediator at follow-up. This causal process represents another type of confounding that is not stable over time.

In all these scenarios, one or more of the parallel trends assumptions may be violated. Even in Panel A, where there is only confounding by a time-invariant factor U, these assumptions could still be violated if the effects of U on the mediator or outcome vary from one time period to the next.

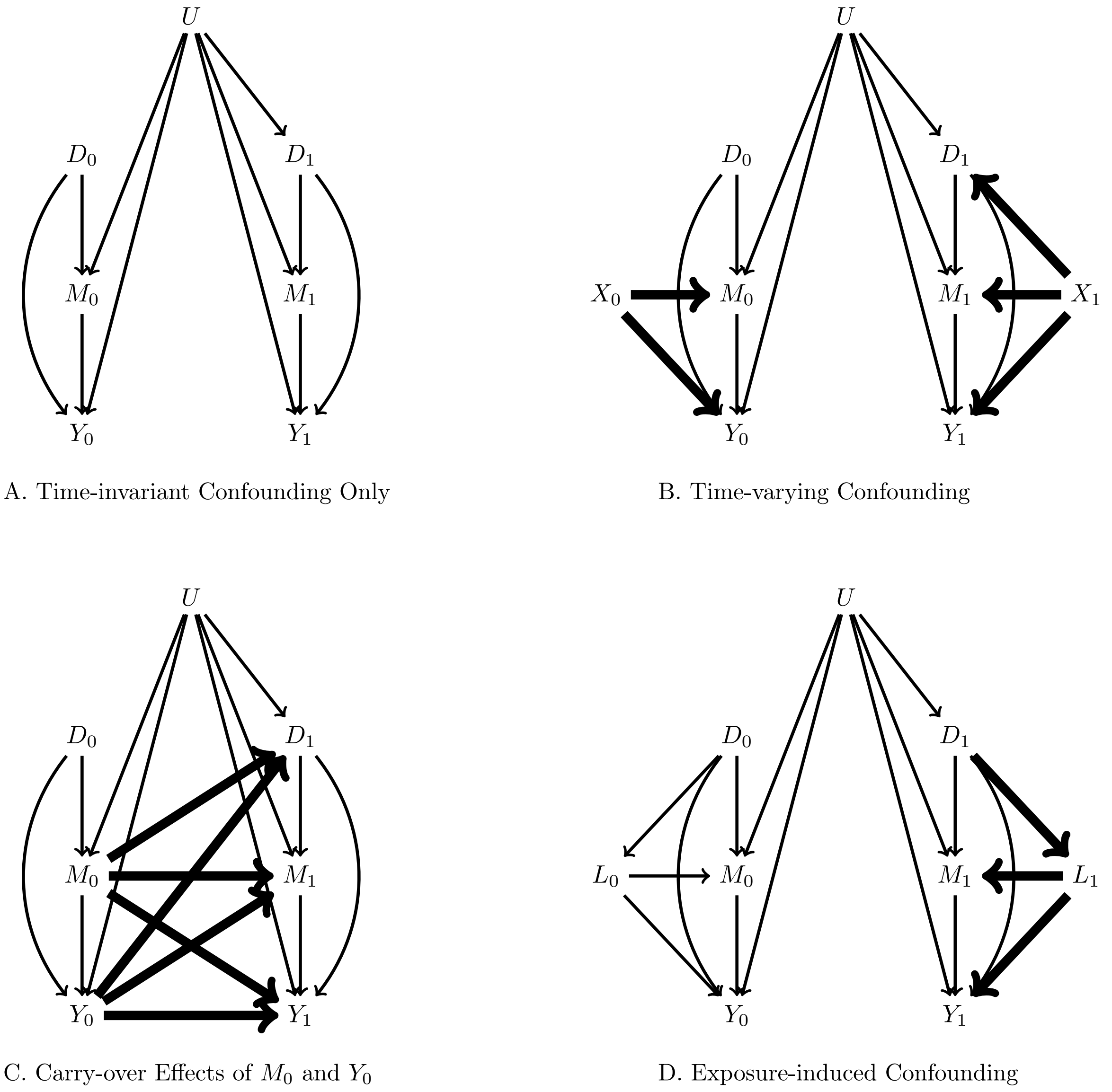
To summarize, assumptions (j.i) to (j.iii) effectively rule out the presence of any carry-over effects, any confounding by time-varying or exposure-induced variables, and any confounding by factors that are time-invariant but exhibit time-varying effects. These are strong assumptions that need not hold in a pre/post design, experimental or otherwise. However, this type of study design can still identify direct and indirect effects if there are no carry-over effects and the only confounding influences are stable over time, regardless of whether these influences are observed or unobserved.

Assumption (j.iv). The assumption of invariant effects requires that the impact of the mediator on the outcome remains constant over time and does not interact with the exposure. These conditions can be formally expressed as follows:

$$Y_0\left(0,m\right) - Y_0\left(0,m^*\right) = Y_1\left(d,m\right) - Y_1\left(d,m^*\right). \tag{7.35}$$

Substantively, this assumption requires that the effect of the mediator on the outcome is identical at baseline and at follow-up, regardless of any exposure received between these periods. It would be violated if the effect of the mediator changes over time or if the exposure and mediator interact to influence the outcome.

If assumptions (j.i) to (j.iv) are satisfied, along with a standard consistency assumption, the natural direct and indirect effects of exposure on the outcome can be estimated using a DiD approach with data from a pre/post design. The estimation process involves two steps. First, the difference in the mediator is regressed on the difference in the exposure using a linear model, which can be expressed as follows:



A. Time-invariant Confounding Only

B. Time-varying Confounding

C. Carry-over Effects of M_0 and Y_0

D. Exposure-induced Confounding

Figure 7.7: Graphical Mediation Models Depicting Different Types of Causal Relations in a Pre/Post Design.

Note: D_t denotes the exposure at time $t \in \{0, 1\}$, where $D_0 = 0$ for everyone because nobody is exposed to treatment at baseline. M_t and and Y_t denote the mediator and outcome, respectively, measured at time t. X_t denotes a time-varying confounder, while L_t denotes a time-varying confounder that is also exposure-induced.

$$\mathbb{E}\left[\Delta M|\Delta D\right] = \theta_0 + \theta_1 \Delta D. \tag{7.36}$$

Second, the difference in the outcome is then regressed on differences in both the exposure and the mediator using another linear model, which has the following form:

$$\mathbb{E}\left[\Delta Y|\Delta D, \Delta M\right] = \lambda_0 + \lambda_1 \Delta D + \lambda_1 \Delta M. \tag{7.37}$$

After fitting these models by least squares, estimates for the natural direct and indirect effects are given by $\widehat{NDE}\left(d, d^*\right)^{DiD} = \hat{\lambda}_1 \left(d - d^*\right)$ and $\widehat{NIE}\left(d, d^*\right)^{DiD} = \hat{\theta}_1 \hat{\lambda}_1 \left(d - d^*\right)$, respectively. Because assumption (j.iv) precludes any exposure-mediator interaction, the $NDE\left(d, d^*\right)$ and $CDE\left(d, d^*, m\right)$ are equivalent in this scenario, and thus $\hat{\lambda}_1 \left(d - d^*\right)$ also serves as a DiD estimator for the controlled direct effect. To compute inferential statistics, we can again use the nonparametric bootstrap.

The consistency of these estimators, and the validity of any inferential statistics derived from the bootstrap, both hinge on the assumptions of parallel trends and invariant effects. Should any of these assumptions fail to hold, then the DiD estimators outlined previously may be biased and inconsistent. Although the assumptions of parallel trends and invariant effects are restrictive and cannot be empirically verified, they do not preclude the presence of time-invariant confounding. Even with this type of confounding, the DiD approach can still yield consistent results, provided its other motivating assumptions are met.

To illustrate how these methods might be applied in practice, we can again revisit the JOBSII study (Vinokur et al. 1995; Vinokur and Schul 1997; Imai et al. 2010a). As outlined previously, this study evaluated the effect of a job training workshop on subsequent employment among a sample of participants who were initially unemployed. It also examined whether this effect was mediated by job search self-efficacy. The study employed a standard experimental design, where participants were randomly assigned to the workshop or a control group, with measures of job search self-efficacy and subsequent employment collected after the workshop concluded. In our prior analyses, detailed in Chapter 3, we relied on strong assumptions about the absence of unobserved and exposure-induced confounding to assess the mediating role of self-efficacy. However, these assumptions were not entirely supported by the original study's design, highlighting the need for other approaches.

Suppose that we redesigned the JOBSII study to include measurements of job search self-efficacy taken both before and after participants were randomly assigned to the workshop or a control group. In this scenario, M_t would represent a measurement of job search self-efficacy taken at time $t \in \{0, 1\}$, where $t = 0$ denotes the baseline measurement before randomization and $t = 1$ denotes the follow-up measurement. Additionally, let Y_t denote an indicator of employment status at time t. Because everyone recruited into the JOBSII study was initially unemployed, Y_0 is coded 0 for all participants at baseline, while Y_1 is coded 1 if a participant had secured employment at follow-up and 0 otherwise. Similarly, D_t indicates participation in the job training workshop, where D_0 is uniformly coded 0 at baseline, since no one had yet participated in the workshop, and D_1 is coded 1 for those assigned to the workshop and 0 for those in the control group.

With this experimental pre/post design, the average total effect of the workshop on employment can be identified and consistently estimated by comparing the employment rates at follow-up between participants assigned to the workshop and those in the control group. Furthermore, the natural direct and indirect effects of the workshop could also be identified and consistently estimated, provided that the following conditions are met: there must be parallel trends in employment and job search self-efficacy with respect to workshop participation, parallel trends in employment with respect to self-efficacy, and invariant effects of self-efficacy on employment.

The assumptions of parallel trends in employment and self-efficacy with respect to workshop participation are met by design in the JOBSII study, as the exposure is randomly assigned. However, the assumption of parallel trends in employment with respect to self-efficacy is not guaranteed, even in the redesigned experiment. This assumption could be violated if the effect of self-efficacy on employment is confounded by time-varying factors or by stable characteristics with time-varying effects. It could also be violated if self-efficacy at baseline has carry-over effects on self-efficacy or employment at follow-up. In addition, the assumption of invariant effects could fail if workshop participation interacts with self-efficacy to influence employment at follow-up. Although these assumptions are strong and unverifiable, they do not necessitate the absence of time-invariant confounding for the relationship between self-efficacy and employment. Essentially, equipping the JOBSII experiment with a pre/post design allows us to estimate natural direct and indirect effects with a DiD approach, which circumvents any bias due to time-invariant forms of confounding by observed or unobserved factors.

To implement this approach, we would begin by regressing the change in levels of job search self-efficacy, $\Delta M = M_1 - M_0$, on the change in exposure, $\Delta D = D_1 - D_0$ from baseline to follow-up. We would then regress the change in employment status, denoted by $\Delta Y = Y_1 - Y_0$, on both ΔD and ΔM, which reflect the differences in exposure and self-efficacy over time. The coefficient on ΔD in the outcome regression provides an estimate of the natural direct effect, while the product of the coefficient on ΔD in the mediator regression and the coefficient on ΔM from the outcome regression yields an estimate of the natural indirect effect. Confidence intervals and hypothesis tests for these effects could be constructed using the nonparametric bootstrap.

Using a pre/post design and DiD methods for mediation analysis obviates the need for assumptions about the absence of time-invariant confounding. Nevertheless, the validity of this approach still hinges on other stringent assumptions that may not be satisfied in many applications. Indeed, these assumptions can be just as onerous as those required of mediation analyses based on a standard experiment or an observational study that lacks repeated measurements. While the DiD approach offers flexibility and can be adapted to relax some of the assumptions outlined previously, such modifications often require shifting focus to different estimands, typically defined over narrower subpopulations, or they rely on other identification conditions whose substantive implications can be difficult to untangle (e.g., Blackwell et al. 2024; Holm and Breen 2023; Huber et al. 2022). Pre/post designs and DiD estimation are powerful methods for addressing certain types of unobserved confounding, but they are no panacea. As with any approach to analyzing causal mediation, they must be applied with caution and prudence in an effort to bolster the validity of findings.

7.4 Summary

In this chapter, we introduced methods for analyzing causal mediation using experimental and quasi-experimental designs, which depend on less stringent assumptions regarding the absence of unobserved confounding compared to non-experimental methods. We explored a variety of experimental designs that utilize random assignment of both the exposure and a mediator of interest, including joint, parallel, and multi-arm randomization designs. We also discussed several quasi-experimental approaches, where the exposure is randomized directly but the mediator is either manipulated indirectly or measured before and after the exposure without manipulation. These methods, including IV and DiD models for direct and indirect effects, also facilitate mediation analyses under relaxed assumptions about unobserved confounding.

Despite their advantages, both experimental and quasi-experimental methods for investigating causal mediation suffer from noteworthy limitations. Experimental approaches are often hampered by the logistical and practical difficulties associated with directly manipulating complex social conditions or psychological constructs. Quasi-experimental methods circumvent some of these challenges but introduce others, since they typically invoke alternative assumptions, such as those requiring the absence of exposure-mediator interaction, that are also problematic in many social science applications, even if they do not preclude confounding by unobserved variables.

Mediation analysis is challenging, but it is not hopeless (Green et al. 2010). The most credible inferences are derived from a preponderance of evidence obtained from different designs and under different sets of assumptions. Experimental and quasi-experimental methods, together with rigorous observational analyses, can collectively contribute to this body of evidence and generate robust inferences.

Experimental and quasi-experimental techniques for analyzing causal mediation are still in their infancy. We focused on a subset of these methods that are most relevant for typical applications in the social sciences, while omitting others due to their complexity, idiosyncrasy, or impracticality. These include crossover (Imai et al. 2013; Pearl 2001; Robins and Greenland 1992) and path-severing experimental designs (Glynn 2021; Pearl 2001), as well as IV methods that instrument both the mediator and outcome (Burgess et al. 2015; Dippel et al. 2017; Frolich and Huber 2017; Jun et al. 2016) or preclude a direct effect of treatment on the outcome (Sobel 2008). We also elided DiD methods targeting effects among specific subpopulations (Blackwell et al. 2024; Holm and Breen 2023; Huber et al. 2022). For those interested in these more complex topics, we recommend consulting the specialized literature cited above.

Chapter 8

The Future of Causal Mediation Analysis

This book provided a comprehensive introduction to causal mediation analysis, with a focus on its application to social science research. We began by outlining the foundational concepts, essential principles, and precise notation that underpin mediation analysis, including the counterfactual framework, potential outcomes, and causal graphs. Our exploration then progressed to analyses of causal mediation involving a single mediator, which are often complicated by baseline and exposure-induced confounding. We introduced a range of methodological approaches—from conventional linear models to more complex simulation methods and weighting techniques—to estimate direct and indirect effects in these settings. As the discussion advanced further, our focus shifted to analyses of multiple mediators. In these applications, we demonstrated how to decompose total effects into multivariate direct and indirect effects, and also into path-specific effects, using flexible regression-imputation and weighting methods.

In subsequent chapters, we introduced robust estimation methods and explored the rapidly advancing field of machine learning, which holds significant promise for addressing the challenge of model misspecification and for enhancing the accuracy of mediation analyses in the social sciences. We also examined the design and analysis of experimental and quasi-experimental studies specifically tailored for analyses of mediation. As part of these discussions, we introduced a set of innovative experimental designs that can effectively mitigate various forms of unobserved confounding and bolster the credibility of causal inferences about direct and indirect effects.

Our synthesis lays a strong foundation for future inquiry and analysis of causal mediation in the social sciences. Over the past several decades, however, methods for analyzing causal mediation have rapidly proliferated and evolved, and it is likely that new developments will continue to emerge well into the future. Throughout this book, we have noted various limitations of the methods discussed and identified other areas ripe for ongoing methodological development, all highlighted in the concluding sections of each chapter. In this final chapter, we will revisit and highlight some of these limitations and open areas of inquiry, charting potential future directions that could expand and enrich the application of causal mediation analysis in social science research.

Moreover, in the interest of clarity and focus, our discussion in previous chapters necessarily excluded certain topics and approaches. These omissions encompassed more sophisticated estimands, such as conditional direct and indirect effects, which can be used to examine "moderated mediation" (Hayes 2017), and alternative decompositions that further partition effects into components capturing different types of mediation and interaction operating together. In addition, we omitted discussions on direct and indirect effects defined using ratios, rather than differences, between potential outcomes, as well as methods for addressing measurement error in mediation analyses. And while we presented a range of experimental and quasi-experimental methods for analyzing mediation, we did not explore more complex or idiosyncratic approaches like crossover designs or instrumental variable methods that utilize separate instruments for both the exposure and mediator. As we look to the future, these areas, among others, offer promising avenues for further research and development, presenting opportunities to enhance and broaden the methods prioritized in previous chapters.

8.1 Alternative Estimands

In this book, we concentrated on identifying and estimating direct and indirect effects for a single target population. However, researchers often aim to investigate these effects across specific subpopulations and examine variability in the causal processes of interest among these groups, which is commonly referred to as moderated mediation. Such analyses typically employ estimands similar to those discussed throughout this book, only now conditioned upon subgroup membership. In many applications, moderated mediation can be analyzed simply by stratifying the data into subsamples, applying the methods we have covered previously to each, and comparing the results across groups. In other instances–for example, when subgroups are defined in terms of variables that are continuous, high-dimensional, or may be influenced by the exposure–analyses of moderated mediation become more complex and demand an alternative approach. While some methods for analyzing differences in mediation across distinct populations have already been developed, as in the specialized literature on conditional direct and indirect effects (Vansteelandt and VanderWeele 2012; VanderWeele 2015), the ubiquity of subgroup heterogeneity in the social sciences underscores the need for future research to develop a comprehensive set of causal estimands and estimation procedures specifically designed for exploring moderated mediation. Following recent work on machine learning for analyzing effect heterogeneity (Wager and Athey 2018), we view algorithmic methods for inductively discovering subgroup differences in direct and indirect effects as an especially promising direction for future development.

Relatedly, researchers in the social sciences often aim to understand how mediation and interaction jointly contribute to a causal effect. Although mediated interaction shares some surface similarities with moderated mediation, it is conceptually distinct; moderated mediation explores variations in direct and indirect effects across observed subpopulations, whereas mediated interaction examines the combined influence of the exposure and mediator on the outcome, assessing whether part of the total effect results from mediation and interaction operating together. In previous chapters, we concentrated on two-way decompositions of total effects into direct and indirect components, which did not specifically isolate the contribution of effect interaction. However, more sophisticated approaches have been developed for certain applications (VanderWeele 2013, 2014; Wodtke and Zhou 2020). These approaches employ three- and even four-way decompositions to further partition total effects into components involving pure mediation, pure interaction, and mediated interaction, but methods for estimating these components remain somewhat limited. Given the importance of mediated interaction in many social science theories, which often posit that mediation and interaction operate simultaneously to connect the exposure to an outcome, future research should continue to refine and advance methods for analyzing these types of complex causal processes.

This book also elided discussions of other, more complicated decompositions, such as those based on relative risks or odds ratios, instead of differences in means. While estimands defined in terms of mean differences are generally more intuitive and relevant for decision-making and policy, there are several scenarios in the social sciences where relative risks or odds ratios are more appropriate. For example, in studies of rare events, relative risks are often more stable and easier to interpret than risk differences. Similarly, in case-control studies where the marginal distribution of the outcome is fixed by design, odds ratios remain identifiable and can be consistently estimated, unlike effects measured by mean differences or relative risks, which cannot be directly estimated from the study data due to the outcome-dependent sampling strategy. Although several effect decompositions that capture causal mediation have been developed using relative risks and odds ratios (Valeri and VanderWeele 2013; VanderWeele and Vansteelandt 2010; Wang and Albert 2012), these could benefit from further extensions to better address exposure-induced confounding, incorporate multiple mediators, and facilitate robust estimation techniques.

8.2 Longitudinal Analyses

Throughout this book, many of our empirical analyses have relied on longitudinal data, which is important for maintaining an appropriate causal sequence among the variables of interest. By following individuals over time and measuring the confounders before the exposure, the exposure before the mediator, and the mediator before the outcome, researchers can ensure the proper temporal ordering of these variables. Without longitudinal data, causal mediation analysis becomes more challenging, as contemporaneous measurements of the exposure, mediator, and outcome can undermine the validity of key identification assumptions. For example, problems arise if the outcome influences the mediator, or the mediator influences the exposure, rather than the other way around. Longitudinal measurements help to mitigate these problems by clarifying the temporal priority among variables and reducing the likelihood of reverse or simultaneous causality, which, if present, would violate the assumptions of independence between the potential outcomes and observed data that are needed to identify direct and indirect effects.

Despite the frequent use of longitudinal data, our discussion largely bypassed methods for analyzing causal mediation with repeated, time-varying measurements of the exposure, mediator, and/or outcome, with a few exceptions such as pre/post experimental designs. Instead, we focused mainly on direct and indirect effects of a single, point-in-time exposure on an outcome measured at the end of follow-up, as mediated by a variable measured at one intermediate period. However, many variables of interest to social scientists change over time, and understanding how the cumulative effects of a time-varying exposure are transmitted through successive time-varying measurements of a mediator poses significant challenges. Such analyses are typically complicated by the presence of time-varying confounders, which may be influenced by prior exposures and mediators, making it difficult to identify longitudinal variants of natural direct and indirect effects. While interventional analogues to these effects, as well as longitudinal variants of controlled direct and eliminated effects, can be identified in the presence of time-varying confounders, consistently estimating them is difficult and often requires stringent modeling assumptions or unstable weighting procedures. Building on foundational work and recent advances in this area (VanderWeele 2009b; VanderWeele and Tchetgen Tchetgen 2017; Tai et al. 2023), future research should strive to develop simpler and more robust estimation methods for longitudinal mediation effects.

As discussed in Chapter 7, longitudinal data and repeated measurements of an exposure, mediators, and outcome can also help to mitigate bias from certain forms of unobserved confounding in some situations. We explored the use of difference-in-difference (DiD) methods to analyze causal mediation in a simple two-time period setting, as part of a pre/post experimental design. DiD methods, along with their close relative, two-way fixed effects models, are used extensively in the social sciences to estimate total effects of an exposure on an outcome, while controlling for time-invariant confounding (Abadie 2005; Allison 2009; Halaby 2004). And yet, despite the many recent advances that have improved their robustness and flexibility (Athey et al. 2021; Callaway and Sant'Anna 2021; Xu 2017), these methods have not been thoroughly adapted for causal mediation analyses. Future research should aim to further refine and extend DiD and fixed effects models for longitudinal data to better capture different types of direct and indirect effects.

Finally, our discussion did not extend to methods for analyzing causal mediation with survival or time-to-event outcomes, which often involve longitudinal data as well. These types of duration outcomes are common in social science research—for example, sociologists frequently study the effects of different social conditions on human survival, time to a first childbirth among women, or time to marital dissolution among married couples (Heaton 1991; Rendall et al. 2011; Wodtke 2013). Mediation analysis with duration outcomes presents some unique challenges. It often involves time-varying measurements of the exposure, mediators, and confounders. Moreover, it is also often complicated by issues like censoring, where the event of interest (e.g., death) may not occur for some individuals by the end of the study, and nonlinearity, where the outcome relates to its causes in more complex ways. Methods for analyzing mediation with duration outcomes have progressed in recent years, including new approaches based on accelerated failure time models (Tein and MacKinnon 2003; VanderWeele 2011a), additive hazard models (Lange and Hansen 2011), and proportional hazard models fit using inverse probability weights (VanderWeele 2015; Vansteelandt et al. 2012). However, further development is necessary to enhance the flexibility and robustness of these approaches, to accommodate multiple mediators or time-varying measures of the exposure, mediator, and confounders, and to assess the sensitivity of estimates to unobserved confounding or informative censoring.

8.3 Complex Sample Designs

Complex sample designs refer to respondent sampling strategies that deviate from simple random sampling, where every possible sample of n respondents from a target population has an equal probability of selection. Modern surveys frequently employ complex random samples in order to enhance efficiency, reduce the costs of data collection, and/or increase the representation of certain subpopulations. However, while complex samples offer many advantages, they also introduce some additional complications for drawing valid inferences. These designs often involve stratification, clustering, and unequal probabilities of sample selection. Stratified sampling divides the population into distinct subgroups and then randomly selects individuals from within each of them. Cluster sampling groups individuals into clusters and then randomly selects entire clusters for inclusion in the study. Stratification and clustering, among other features of complex sample designs, often lead to unequal probabilities of sample selection, where certain members of the target population are more (or less) likely to be selected than others. This can generate samples that are not representative of the target population, with some groups under-represented and others over-represented, possibly necessitating the use of sampling weights to produce unbiased and consistent estimates.

In analyses of causal mediation, these features of the sample design often require special consideration to generate accurate point estimates of direct and indirect effects, and to appropriately quantify their random variability due to sampling error. To account for unequal probabilities of selection, inverse probability of sample selection weights can be used to ensure that estimation procedures remain consistent and unbiased (Lumley 2011). Although our previous discussions did not specifically address adjustments for complex sample designs, all of the estimation procedures outlined in earlier chapters can accommodate sampling weights that correct for unequal probabilities of sample selection. For example, when estimating direct and indirect effects using linear models, analysts can simply compute weighted least squares estimates for these models, using the inverse probabilities of sample selection as weights. The coefficients obtained can then be used to construct direct and indirect effect estimates, exactly as outlined in previous chapters. Similarly, for simulation and regression imputation estimators, all models can be fit using sampling weights, and then the simulated or imputed potential outcomes are averaged together using these weights as well. Estimation procedures based on inverse probability weighting can also be adapted to incorporate sampling weights. To this end, models for the exposure and/or mediator are initially fit using the sampling weights. Then, the inverse probability weights derived from these models should be multiplied by the sampling weights before they are used to estimate outcome means, and by extension, to compute direct and indirect effects.

Apart from the additional complexities posed by unequal probabilities of sample selection, stratification and clustering can also influence the random variability in an estimator due to sampling error, leading to what are known as *design effects* (Lumley 2011). Ignoring design effects can result in invalid inferences, even if the estimation procedures remain unbiased and consistent, because they distort the width of confidence intervals and the size of p-values. Throughout this book, we focused mainly on a conventional implementation of the bootstrap, where individuals are resampled with replacement from the observed data to quantify sampling variability. However, without modifications, this method does not account for design effects that arise from stratification or clustering, potentially leading to confidence intervals that are too narrow—or occasionally too wide—as well as p-values that may also be inaccurate. Fortunately, the bootstrap can be adapted to accommodate complex sample designs. For example, to adjust for clustering, the cluster bootstrap can be employed, which involves resampling entire groups of respondents with replacement (Cameron et al. 2008; Efron and Tibshirani 1994). To adjust for stratification, individuals can be resampled within each design stratum before being combined to form the final bootstrap sample. For designs that incorporate stratification, clustering, and unequal probabilities of selection altogether, bootstrap replicate weights can be constructed by resampling the observed data in a manner that mirrors the survey's original sampling strategy and then adjusting the corresponding sampling weights based on the frequency with which a cluster or individual appears in the bootstrap sample (Rao and Wu 1988). These modifications ensure that the bootstrap accurately accounts for design effects in mediation analyses of complex samples, and they can generally be implemented without too much difficulty in most software packages, including R and Stata.

Complex sample designs can also yield data with a hierarchical structure, where individuals are nested within clusters, which are themselves nested within larger strata. For example, in education research, sample surveys often produce data where students are nested within classrooms, which are nested within schools, and these in turn may be nested within higher-order administrative units, such as districts or states. With hierarchical data, social scientists might aim not only to adjust for design effects stemming from complex sampling but also to explore how causal effects vary across these nested clusters or strata. How does the impact of an instructional intervention on students differ across classrooms or schools? What factors might explain variation in these treatment effects? Although multilevel models are increasingly employed to investigate these types of questions (Gelman and Hill 2006; Raudenbush and Bryk 2002; Raudenbush and Schwartz 2020), their application to studying how causal processes vary across different contexts is still nascent (Qin and Hong 2017; Qin et al. 2019, 2021). Future research should focus on expanding the use of multilevel models to analyze variability in direct and indirect effects in order to enhance analyses of data from complex samples and to enrich our understanding of contextual variability in causal processes.

8.4 Measurement Error

Measurement error arises when the variables used in an analysis fail to accurately represent the true values of the concepts they are intended to measure. This discrepancy between observed and true values may stem from many different sources, including the use of imprecise measurement instruments, respondent misreporting, or data entry errors, among other possibilities. All these challenges commonly afflict social science research, and thus measurement error is endemic.

In the context of causal mediation analysis, measurement error can complicate estimation and inference by introducing several types of bias. Specifically, inaccurate measurements of key variables, including the exposure, mediator, confounders, or outcome, can lead to biased estimates of both direct and indirect effects. For example, measurement error in a mediator typically biases estimates of indirect effects toward zero and inflates estimates of direct effects away from zero, although other biases may arise depending on the nature and magnitude of the error (VanderWeele et al. 2012). Similarly, mismeasurement of an exposure generally leads to attenuation bias in estimates of direct effects toward zero, while its impact on indirect effect estimates can vary, potentially leading to under or overestimation, again depending on the type of measurement error (Cheng et al. 2023; Jiang and VanderWeele 2019).

Measurement errors in the confounders can introduce additional biases, even in the absence of any confounding factors that are entirely unobserved. In this case, adjusting for a mismeasured confounder may only partially eliminate confounding bias in estimates of direct and indirect effects. In addition, certain types of measurement error in the outcome can lead to bias in estimates of direct and indirect effects as well (Jiang and VanderWeele 2015), and even when such errors do not induce bias, they can still reduce precision and inflate the sampling variance of effect estimates. Altogether, these different types of measurement error can seriously undermine the validity of conclusions drawn from mediation analyses.

When estimating direct and indirect effects using simple parametric models, such as linear or logistic regressions without exposure-mediator interactions, adjustments for measurement error are well-established and straightforward to implement (Bollen 1989; Cheng et al. 2023; le Cessie et al. 2012; VanderWeele et al. 2012). However, as we have discussed throughout this book, simple parametric models often fail to accurately capture the complex relationships among variables in social science analyses of mediation. In applications with more sophisticated models or exposure-mediator interactions, adjusting for measurement error becomes considerably more difficult. Although methods are available to address these complexities in certain situations (Valeri et al. 2014), the options available for implementing such corrections are still limited, and techniques for addressing measurement error in analyses that employ robust estimation methods or machine learning remain largely undeveloped. Future research should further explore methods to adjust for measurement error across a wider range of applications and within a diverse array of more versatile models.

8.5 Robust Estimation Methods

In Chapter 6, we explored the application of robust estimation methods using machine learning to analyses of causal mediation. This approach is particularly powerful as it can mitigate bias that might otherwise arise from model misspecification while maintaining a high level of precision in the resulting effect estimates. Unlike more conventional methods, where protecting against misspecification bias often involves increasing model complexity—and consequently the sampling variability of estimates—robust estimation techniques utilizing machine learning algorithms can often avoid biases stemming from model misspecification without compromising very much on precision.

Despite the advantages of robust estimation using machine learning methods, this approach typically requires discrete exposures and/or mediators for effective implementation, and most of the existing research on robust estimation for causal effects has sidestepped applications involving continuous or many-valued variables. This limitation poses significant challenges for analyses of causal mediation, where continuous or many-valued exposures and mediators are frequently encountered, particularly in observational studies. For example, throughout this book, we examined the effects of a discrete, binary exposure—attending college versus not—on later life depression, but researchers might also be interested in exploring the relationship between mental health and many different levels of education, such as years of completed schooling. However, because years of schooling involves a wide range of values and its relationship with depression is best represented by a dose-response curve, applying robust estimation methods to analyze these effects is much more challenging.

While several recent advances in robust estimation methods for continuous exposures have improved the analysis of total effects (Diaz and van der Laan 2013; Kennedy et al. 2017), these techniques have yet to be extended to causal mediation analyses that aim to decompose total effects into direct and indirect components. Future research should focus on adapting these methods to handle mediation analyses involving continuous or many-valued exposures and mediators.

Another promising direction for future development involves a class of models known as graphical normalizing flows (GNFs), which utilize deep neural networks to model the entire system of relationships among variables in a causal graph (Balgi et al. 2024; Javaloy et al. 2023; Wehenkel and Louppe 2021). Unlike traditional methods, GNFs model the full joint distribution of the data based on a causal graph specified by the analyst, all without invoking any stringent assumptions about its functional form. This enables the construction of estimates for any causal estimand identified from the graph, including total effects, direct and indirect effects, and path-specific effects, that are essentially nonparametric in nature. Moreover, a key advantage of GNFs is that they can accommodate both continuous and discrete variables (Balgi et al. 2024). Continuing to explore and improve the utility of normalizing flows therefore presents another encouraging direction for future research on causal mediation analysis (Zhou and Wodtke 2024).

8.6 Experimental and Quasi-experimental Methods

The development of experimental and quasi-experimental methods for analyzing causal mediation is another area that presents considerable potential for growth. We discussed several such approaches in Chapter 7, including parallel experimental designs, multi-arm experimental designs, encouragement designs analyzed with instrumental variable (IV) methods, and pre/post designs analyzed with difference-in-difference (DiD) methods. However, these approaches are still in their infancy and have not been widely used in practice. Future research should concentrate on refining these methods to enhance their robustness and practical utility, while also adapting them to explore a wider range of causal processes.

At present, the variety of experimental designs available for identifying and estimating direct and indirect effects is still quite limited, particularly when compared against the experimental methods available for analyzing total effects. Expanding the range of these approaches will surely help to improve the reliability of evidence about causal mediation generated from social science research. Additional work is needed to develop more comprehensive power and sample size calculations for mediation analyses, though recent advances have made important progress in this area (Qin 2024; Schoemann et al. 2017; VanderWeele 2020; Zhang 2014).

As the use of experimental study designs in research on causal mediation becomes more widespread, it will be essential to have clear guidance on the amount of data required to precisely estimate different types of direct and indirect effects and to conduct sufficiently sensitive hypothesis tests.

Randomized encouragement designs and IV methods that utilize separate instruments for the mediator and the exposure are also promising avenues for further development and refinement (Burgess et al. 2015; Frolich and Huber 2017; Jun et al. 2016). Similarly, DiD and two-way fixed effects models that address staggered treatment adoption, exposure-induced confounding, and other complexities in analyses of mediation are ripe for new advances (Blackwell et al. 2022, 2024; Blackwell and Yamauchi 2021). In general, we anticipate that experimental and quasi-experimental methods for analyzing mediation will evolve rapidly in the coming years.

8.7 Conclusion

As we conclude our exploration of causal mediation analysis, it is clear that research and methods in this area have made tremendous advances over the past several decades, and yet, they still continue to present abundant opportunities for further development. The issues we have discussed—from robust estimation techniques and innovative experimental designs to the challenges posed by measurement error and complex sample designs—highlight both the progress already achieved and the pitfalls that remain. Future research is poised not only to continue this progress but also to transform lingering challenges into groundbreaking innovations. By continuing to build, refine, and improve methods for analyzing mediation, we can enhance their utility and accuracy across applications in the social sciences and beyond. These innovations will deepen our understanding of the mechanisms through which causes affect outcomes and provide more reliable evidence to inform theory, policy, and practice. As we push the boundaries of existing methods and cultivate deeper knowledge of causal processes in the social world, we move closer to fully unraveling the intricate web of factors that shape our lives, thoughts, and behaviors.

Appendix A

Nonparametric Identification of the Average Total Effect

If assumptions (a.i) to (a.iii) are all satisfied, the $ATE\left(d, d^*\right)$ can be nonparametrically identified as follows:

$$\begin{aligned}
ATE\left(d, d^*\right) &= \mathbb{E}\left[Y\left(d\right) - Y\left(d^*\right)\right] && \\
&= \mathbb{E}\left[Y\left(d\right)\right] - \mathbb{E}\left[Y\left(d^*\right)\right] && \textit{by properties of } \mathbb{E}\left[\cdot\right] \\
&= \mathbb{E}_C\left[\mathbb{E}\left[Y\left(d\right)|C\right] - \mathbb{E}\left[Y\left(d^*\right)|C\right]\right] && \textit{by iterated expectations} \\
&= \mathbb{E}_C\left[\mathbb{E}\left[Y\left(d\right)|C, d\right] - \mathbb{E}\left[Y\left(d^*\right)|C, d^*\right]\right] && \textit{by (a.i) and (a.ii)} \\
&= \mathbb{E}_C\left[\mathbb{E}\left[Y|C, d\right] - \mathbb{E}\left[Y|C, d^*\right]\right], && \textit{by (a.iii)}
\end{aligned}$$

where $\mathbb{E}_C\left[\cdot\right]$ denotes an expected value, or average, taken over the distribution C. We arrived at the final equality of this derivation by starting with the formal definition of the $ATE\left(d, d^*\right)$ in terms of potential outcomes and then by invoking several properties of expected values together with assumptions (a.i) to (a.iii). Specifically, one of the properties of expected values on which we rely in this derivation is that $\mathbb{E}\left[A - B\right] = \mathbb{E}\left[A\right] - \mathbb{E}\left[B\right]$, which means that the average of the difference between any two random variables, A and B, is equal to the difference between their averages. We also rely on the law of iterated expectations, which states that $\mathbb{E}\left[A\right] = \mathbb{E}_B\left[\mathbb{E}\left[A|B\right]\right]$, or in other words, that the expected value of one random variable A is equal to the average of its conditional expected values given another random variable B taken over the distribution of B.

The conditional independence assumption (a.i) allows us to make $\mathbb{E}\left[Y\left(d\right)|C\right]$ also conditional on D because, if $Y\left(d\right)$ is statistically independent of D given C, then $\mathbb{E}\left[Y\left(d\right)|C\right] = \mathbb{E}\left[Y\left(d\right)|C, D\right]$. The positivity assumption (a.ii) allows us to set D at any of its possible values in $\mathbb{E}\left[Y\left(d\right)|C, D\right]$ while ensuring that this expected value remains well-defined. And the consistency assumption (a.iii) allows us to equate $\mathbb{E}\left[Y\left(d\right)|C, D\right]$ with $\mathbb{E}\left[Y|C, d\right]$ and $\mathbb{E}\left[Y\left(d^*\right)|C, d^*\right]$ with $\mathbb{E}\left[Y|C, d^*\right]$ because, under this assumption, Y is equal to $Y\left(d\right)$ among individuals for whom $D = d$, and Y is equal to $Y\left(d^*\right)$ among individuals for whom $D = d^*$. By invoking these properties of expected values and the nonparametric identification assumptions (a.i) to (a.iii), we are therefore able to equate the average total effect with a function of observable data rather than counterfactual quantities. This function from the final equality in the derivation above, $\mathbb{E}_C\left[\mathbb{E}\left[Y|C, d\right] - \mathbb{E}\left[Y|C, d^*\right]\right]$, can also be written as follows:

$$ATE\left(d,d^{*}\right)=\sum_{c}\left(\mathbb{E}\left[Y|c,d\right]-\mathbb{E}\left[Y|c,d^{*}\right]\right)P\left(c\right),$$

because of the definition of $\mathbb{E}_{C}\left[\mathbb{E}\left[Y|C,d\right]-\mathbb{E}\left[Y|C,d^{*}\right]\right]$ as the probability-weighted average of $\mathbb{E}\left[Y|C,d\right]-\mathbb{E}\left[Y|C,d^{*}\right]$ taken over the distribution of C. This is exactly the same expression given in Equation 3.9 from Section 3.3.1. If C were continuous, the probability-weighted sum in this expression would just be replaced with a density-weighted integral.

Appendix B

Nonparametric Identification of the Controlled Direct Effect

If assumptions (b.i) to (b.iv) are all satisfied, the $CDE\left(d, d^*, m\right)$ can be nonparametrically identified as follows:

$$\begin{aligned}
CDE(d, d^*, m) &= \mathbb{E}\left[Y\left(d, m\right) - Y\left(d^*, m\right)\right] && \\
&= \mathbb{E}\left[Y\left(d, m\right)\right] - \mathbb{E}\left[Y\left(d^*, m\right)\right] && \textit{by properties of } \mathbb{E}\left[\cdot\right] \\
&= \mathbb{E}_C\left[\mathbb{E}\left[Y\left(d, m\right)|C\right] - \mathbb{E}\left[Y\left(d^*, m\right)|C\right]\right] && \textit{by iterated expectations} \\
&= \mathbb{E}_C\left[\mathbb{E}\left[Y\left(d, m\right)|C, d, m\right] - \mathbb{E}\left[Y\left(d^*, m\right)|C, d^*, m\right]\right] && \textit{by (b.i) to (b.iii)} \\
&= \mathbb{E}_C\left[\mathbb{E}\left[Y|C, d, m\right] - \mathbb{E}\left[Y|C, d^*, m\right]\right], && \textit{by (b.iv)}
\end{aligned}$$

where the final equality in this derivation gives an expression for the $CDE(d, d^*, m)$ that is defined only in terms of the observable data $\{Y, C, D, M\}$ and does not impose any functional form restrictions on the probability distribution from which these data are generated. We arrived at this expression by starting with the formal definition of the controlled direct effect and then by invoking assumptions (b.i) to (b.iv) together with the same properties of expected values that we previously outlined in Appendix A. This expression, $\mathbb{E}_C\left[\mathbb{E}\left[Y|C, d, m\right] - \mathbb{E}\left[Y|C, d^*, m\right]\right]$, can also be written as follows:

$$CDE\left(d, d^*, m\right) = \sum_{c}\left(\mathbb{E}\left[Y|c, d^*, m\right] - \mathbb{E}\left[Y|c, d^*, m\right]\right)P\left(c\right),$$

because $\mathbb{E}_C\left[\cdot\right]$ is just an expected value taken over the distribution C. If C were continuous, the probability-weighted sum would be replaced with a density-weighted integral.

Appendix C

Nonparametric Identification of the Natural Direct Effect

Under assumptions (c.i) to (c.vi), the $NDE\left(d, d^*\right)$ can be nonparametrically identified as follows:

$$\begin{aligned}
NDE\left(d, d^*\right) &= \mathbb{E}\left[Y\left(d, M\left(d^*\right)\right) - Y\left(d^*, M\left(d^*\right)\right)\right] \\
&= \mathbb{E}\left[Y\left(d, M\left(d^*\right)\right)\right] - \mathbb{E}\left[Y\left(d^*, M\left(d^*\right)\right)\right] && \textit{by properties of } \mathbb{E}\left[\cdot\right] \\
&= \mathbb{E}_C\left[\mathbb{E}\left[Y\left(d, M\left(d^*\right)\right)|C\right] - \mathbb{E}\left[Y\left(d^*, M\left(d^*\right)\right)|C\right]\right] && \textit{by iterated expectations} \\
&= \mathbb{E}_C[\sum_m (\mathbb{E}\left[Y\left(d, m\right)|C, M\left(d^*\right) = m\right] \\
&\qquad - \mathbb{E}\left[Y\left(d^*, m\right)|C, M\left(d^*\right) = m\right]) P\left(M\left(d^*\right) = m|C\right)] && \textit{by iterated expectations} \\
&= \mathbb{E}_C[\sum_m \left(\mathbb{E}\left[Y\left(d, m\right)|C, d, m\right] - \mathbb{E}\left[Y\left(d^*, m\right)|C, d^*, m\right]\right) \\
&\qquad \times P\left(M\left(d^*\right) = m|C, d^*\right)] && \textit{by (c.i) to (c.v)} \\
&= \mathbb{E}_C[\sum_m \left(\mathbb{E}\left[Y|C, d, m\right] - \mathbb{E}\left[Y|C, d^*, m\right]\right) P\left(m|C, d^*\right)], && \textit{by (c.vi)}
\end{aligned}$$

where the final equality in this derivation gives an expression for the $NDE\left(d, d^*\right)$ that is defined only in terms of the observable data $\{Y, C, D, M\}$ and does not impose any functional form restrictions on the probability distribution from which these data were generated. We arrived at this expression by starting with the formal definition of the natural direct effect in terms of nested and cross-world potential outcomes and then by invoking assumptions (c.i) to (c.vi) together with the same properties of expected values that we previously outlined in Appendix A. It can also be written as follows:

$$NDE\left(d, d^*\right) = \sum_{m,c} \left(\mathbb{E}\left[Y|c, d, m\right] - \mathbb{E}\left[Y|c, d^*, m\right]\right) P\left(m|c, d^*\right) P\left(c\right),$$

because $\mathbb{E}_C\left[\cdot\right]$ is just an expected value taken over the distribution C. If C and M were continuous, the probability-weighted sums would be replaced with density-weighted integrals.

Appendix D

Nonparametric Identification of the Natural Indirect Effect

Under assumptions (c.i) to (c.vi), the $NIE(d, d^*)$ can be nonparametrically identified as follows:

$$\begin{aligned}
NIE(d, d^*) &= \mathbb{E}\left[Y(d, M(d)) - Y(d, M(d^*))\right] \\
&= \mathbb{E}\left[Y(d, M(d))\right] - \mathbb{E}\left[Y(d, M(d^*))\right] && \textit{by properties of } \mathbb{E}[\cdot] \\
&= \mathbb{E}_C\left[\mathbb{E}\left[Y(d, M(d))|C\right] - \mathbb{E}\left[Y(d, M(d^*))|C\right]\right] && \textit{by iterated expectations} \\
&= \mathbb{E}_C[\sum_m \mathbb{E}\left[Y(d, m)|C, M(d) = m\right] P(M(d) = m|C) \\
&\quad - \sum_m \mathbb{E}\left[Y(d, m)|C, M(d^*) = m\right] P(M(d^*) = m|C)] && \textit{by iterated expectations} \\
&= \mathbb{E}_C[\sum_m \mathbb{E}\left[Y(d, m)|C, d, m\right](P(M(d) = m|C, d) \\
&\quad - P(M(d^*) = m|C, d^*))] && \textit{by (c.i) to (c.v)} \\
&= \mathbb{E}_C[\sum_m \mathbb{E}\left[Y|C, d, m\right](P(m|C, d) - P(m|C, d^*))], && \textit{by (c.vi)}
\end{aligned}$$

where the final equality in this derivation gives an expression for the $NIE(d, d^*)$ that is defined only in terms of the observable data $\{Y, C, D, M\}$ and does not impose any functional form restrictions on the probability distribution from which these data were generated. We arrived at this expression by starting with the formal definition of the natural indirect effect in terms of nested and cross-world potential outcomes and then by invoking assumptions (c.i) to (c.vi). We also relied on several properties of expected values, which were previously described in Appendix A. The final equality in this derivation can also be written as follows:

$$NIE(d, d^*) = \sum_{m,c} \mathbb{E}\left[Y|c, d, m\right](P(m|c, d) - P(m|c, d^*))P(c),$$

because $\mathbb{E}_C[\cdot]$ is just an expected value taken over the distribution C. If C and M were continuous, the probability-weighted sums would be replaced with density-weighted integrals.

The nonparametric identification formulas for both the natural direct and indirect effects share an important expression: $\sum_{m,c} \mathbb{E}\left[Y|c, d, m\right] P(m|c, d^*) P(c)$. This quantity is obtained by first computing a set of outcome means given the baseline confounders, level d of the exposure, and level m of the mediator, which is denoted by $\mathbb{E}\left[Y|c,d,m\right]$. These outcome means are then averaged over the distribution of the mediator given the other level of the exposure d^* and the baseline confounders, which is denoted by $P\left(m|c,d^*\right)$. Finally, these quantities are averaged again over the marginal distribution of the baseline confounders, denoted by $P\left(c\right)$. Under the assumptions outlined previously, the result of these calculations is equal to the average of the cross-world potential outcomes that are used to define natural direct and indirect effects, $Y\left(d,M\left(d^*\right)\right)$, where the exposure is set to one level d but the mediator is set to its value that would have arisen under exposure to d^*.

In the nonparametric identification formula for the $NDE\left(d,d^*\right)$, this first quantity is contrasted with another: $\sum_{m,c}\mathbb{E}\left[Y|c,d^*,m\right]P\left(m|c,d^*\right)P\left(c\right)$. This second quantity is equal to the mean of the nested potential outcomes, $Y\left(d^*,M\left(d^*\right)\right)$, under level d^* of the exposure and under the level of the mediator that would follow naturally from this same exposure, $M\left(d^*\right)$, if assumptions (c.i) to (c.vi) are all satisfied. It is also equal to the mean of the conventional potential outcomes, $Y\left(d^*\right)$, if assumptions (a.i) to (a.iii) are satisfied because $\sum_{m,c}\mathbb{E}\left[Y|c,d^*,m\right]P\left(m|c,d^*\right)P\left(c\right)=\sum_{c}\mathbb{E}\left[Y|c,d^*\right]P\left(c\right)=\mathbb{E}\left[Y\left(d^*\right)\right]$, which readers may recognize from the nonparametric identification formula for the $ATE\left(d,d^*\right)$.

In the nonparametric identification formula for the $NIE\left(d,d^*\right)$, on the other hand, $\sum_{m,c}\mathbb{E}\left[Y|c,d,m\right]P\left(m|c,d^*\right)P\left(c\right)$ is contrasted with a different quantity–specifically, it is contrasted with $\sum_{m,c}\mathbb{E}\left[Y|c,d,m\right]P\left(m|c,d\right)P\left(c\right)$. This quantity is equal to the mean of the nested potential outcomes, $Y\left(d,M\left(d\right)\right)$, under level d of the exposure and under the level of the mediator that would follow naturally from this same exposure, $M\left(d\right)$, if assumptions (c.i) to (c.vi) are satisfied. It is also equal to the mean of the conventional potential outcomes, $Y\left(d\right)$, if assumptions (a.i) to (a.iii) are satisfied. This is because $\sum_{m,c}\mathbb{E}\left[Y|c,d,m\right]P\left(m|c,d\right)P\left(c\right)=\sum_{c}\mathbb{E}\left[Y|c,d\right]P\left(c\right)=\mathbb{E}\left[Y\left(d\right)\right]$, which readers may again recognize from the nonparametric identification formula for the $ATE\left(d,d^*\right)$. Thus, to nonparametrically identify natural direct and indirect effects, it suffices to identify the average total effect together with the average of the cross-world world potential outcomes, $\mathbb{E}\left[Y\left(d,M\left(d^*\right)\right)\right]$.

Appendix E

Natural Effects under Linear and Additive Models

In this section, we derive the parametric expressions for the total, direct, and indirect effects of interest under linear and additive models for the mediator and outcome. Consider first the $NDE\left(d, d^{*}\right)$. Given the models described in Equations 3.27 and 3.28, the parametric expression for the $NDE\left(d, d^{*}\right)$ just comes from substituting our linear model for the outcome into the nonparametric identification formula and then simplifying the expression as follows:

$$\begin{aligned}
NDE\left(d, d^{*}\right) &= \sum_{m,c}\left(\mathbb{E}\left[Y|c,d,m\right]-\mathbb{E}\left[Y|c,d^{*},m\right]\right)P\left(m|c,d^{*}\right)P\left(c\right) && \textit{by (c.i) to (c.vi)}\\
&= \sum_{m,c}((\gamma_0+\gamma_1^T c+\gamma_2 d+\gamma_3 m) \\
&\qquad -(\gamma_0+\gamma_1^T c+\gamma_2 d^{*}+\gamma_3 m))P\left(m|c,d^{*}\right)P\left(c\right) && \textit{by model assumption}\\
&= \sum_{m,c}\left(\gamma_2 d-\gamma_2 d^{*}\right)P\left(m|c,d^{*}\right)P\left(c\right) && \textit{by subtraction}\\
&= \gamma_2\left(d-d^{*}\right)\sum_{m,c}P\left(m|c,d^{*}\right)P\left(c\right) && \textit{by distributive law}\\
&= \gamma_2\left(d-d^{*}\right). && \textit{by probability axioms}
\end{aligned}$$

The first line in this derivation just gives the nonparametric identification formula in Equation 3.19 for the natural direct effect. The equality in the second line comes by assumption, namely, that $\mathbb{E}\left[Y|c,d,m\right] = \gamma_0+\gamma_1^T c+\gamma_2 d+\gamma_3 m$, or in other words, that our model for the conditional mean of the outcome in Equation 3.28 is correctly specified. The third equality comes from subtracting terms, the fourth from factoring out constants following the distributive law, and the final equality from the axiom that a sum over a probability distribution must always equal 1.

Next, consider the parametric expression for $NIE\left(d, d^{*}\right)$, which is obtained from a similar procedure. Specifically, the parametric expression for the $NIE\left(d, d^{*}\right)$ is obtained by substituting our linear models for the outcome and the mediator into the nonparametric identification formula and then simplifying as follows:

$$\begin{aligned}
NIE\left(d,d^{*}\right) &= \sum_{m,c} \mathbb{E}\left[Y|c,d,m\right]\left(P\left(m|c,d\right)-P\left(m|c,d^{*}\right)\right)P\left(c\right) && \textit{by (c.i) to (c.vi)}\\
&= \sum_{m,c}\left(\gamma_0+\gamma_1^T c+\gamma_2 d+\gamma_3 m\right)\left(P\left(m|c,d\right)-P\left(m|c,d^{*}\right)\right)P\left(c\right) && \textit{by model assumption}\\
&= \sum_{c}(\gamma_0+\gamma_1^T c+\gamma_2 d+\gamma_3\sum_{m} mP\left(m|c,d\right))\\
&\quad -(\gamma_0+\gamma_1^T c+\gamma_2 d+\gamma_3\sum_{m} mP\left(m|c,d^{*}\right))P\left(c\right) && \textit{by distributive law}\\
&= \gamma_3\sum_{c}(\sum_{m} mP\left(m|c,d\right)-\sum_{m} mP\left(m|c,d^{*}\right))P\left(c\right) && \textit{by subtr. and distr. law}\\
&= \gamma_3\sum_{c}\left(\mathbb{E}\left[M|c,d\right]-\mathbb{E}\left[M|c,d^{*}\right]\right)P\left(c\right) && \textit{by definition of } E\left(\cdot\right)\\
&= \gamma_3\sum_{c}\left(\left(\beta_0+\beta_1^T c+\beta_2 d\right)-\left(\beta_0+\beta_1^T c+\beta_2 d^{*}\right)\right)P\left(c\right) && \textit{by model assumption}\\
&= \beta_2\gamma_3\left(d-d^{*}\right)\sum_{c} P\left(c\right) && \textit{by subtr. and distr. law}\\
&= \beta_2\gamma_3\left(d-d^{*}\right), && \textit{by probability axioms}
\end{aligned}$$

which gives the classic product of coefficients for the indirect effect.

And finally, consider the $CDE\left(d,d^{*},m\right)$. The parametric expression for this effect is also obtained by substituting the model for the outcome into the nonparametric identification formula as follows:

$$\begin{aligned}
CDE\left(d,d^{*},m\right) &= \sum_{c}\left(\mathbb{E}\left[Y|c,d,m\right]-\mathbb{E}\left[Y|c,d^{*},m\right]\right)P\left(c\right) && \textit{by (b.i) to (b.iv)}\\
&= \sum_{c}((\gamma_0+\gamma_1^T c+\gamma_2 d+\gamma_3 m)\\
&\quad -(\gamma_0+\gamma_1^T c+\gamma_2 d^{*}+\gamma_3 m))P\left(c\right) && \textit{by model assumption}\\
&= \sum_{c}\left(\gamma_2 d-\gamma_2 d^{*}\right)P\left(c\right) && \textit{by subtraction}\\
&= \gamma_2\left(d-d^{*}\right)\sum_{c} P\left(c\right) && \textit{by distributive law}\\
&= \gamma_2\left(d-d^{*}\right), && \textit{by probability axioms}
\end{aligned}$$

which is identical to the parametric expression for the $NDE\left(d,d^{*}\right)$ because our assumed models are linear and additive in the predictors.

In sum, these expressions are obtained by substituting the linear and additive models for the mediator and outcome into the nonparametric identification formulas for each targeted estimand, where appropriate, and then simplifying. Under the identification assumptions outlined in Section 3.3 and under the additional assumption of correct model specification, the total, direct, and indirect effects of interest are then equal to simple functions of a few coefficients in these models.

Appendix F

Ratio of Mediator Probability Weighting

In this section, we outline an alternative approach to estimating total, natural direct, and natural indirect effects known as ratio of mediator probability weighting (Hong et al. 2015; Hong 2015). This alternative approach is very similar to the weighting approach outlined in Section 3.5.3, but it requires models for the exposure and the mediator instead of two separate models for the exposure with different sets of predictors.

Ratio of mediator probability weighting is based on an equivalence between two different expressions for the weight used to generate the cross-world pseudosample described in Section 3.5.3. Specifically, the third set of weights from Section 3.5.3, which are given by $w_3 = {}^{P(d^*|C,M)}/_{P(d|C,M)P(d^*|C)}$, can also be expressed as $w_3^{alt} = {}^{P(M|C,d^*)}/_{P(M|C,d)P(d|C)}$. This alternative expression for the weights comes from an application of Bayes' rule, which yields a weight defined in terms an inverse probability of exposure, ${}^{1}/_{P(d|C)}$, multiplied by a ratio of mediator probabilities ${}^{P(M|C,d^*)}/_{P(M|C,d)}$.

The numerator of this ratio, $P(M|C,d^*)$, represents the probability that a sample member experiences their observed mediator M under exposure to d^* conditional on their baseline confounders, while the denominator, $P(M|C,d)$, is the probability that a sample member experiences their observed value of mediator M under exposure to d conditional on the confounders. Among sample members for whom $D = d$, weighting by ${}^{P(M|C,d^*)}/_{P(M|C,d)}$ transforms the distribution of the mediator to resemble its distribution among sample members for whom $D = d^*$. After this transformation is achieved, additionally weighting the subsample for whom $D = d$ by ${}^{1}/_{P(d|C)}$ further transforms the distribution of the confounders so that it mirrors their distribution found in the total sample. Thus, weighting by either w_3 or w_3^{alt} creates the same cross-world pseudosample in which sample members exposed to d appear to have the distribution of C among the total sample and the distribution of M among those exposed to d^*.

This equivalence points toward an alternative set of weighting estimators for total, natural direct, and natural indirect effects that follows nearly the same steps as outlined in Section 3.5.3 but with several small modifications. For this alternative implementation, we would fit a GLM for the mediator given the confounders and the exposure instead of a GLM for the exposure given the confounders and the mediator. We would next use this fitted model to predict the mediator probabilities, $\hat{P}(M|C,d^*)$ and $\hat{P}(M|C,d)$, in the numerator and denominator of w_3^{alt}. Estimators for the total, natural direct, and natural indirect effects of interest would then be constructed using the same expressions as in Equation 3.40 but with $\hat{w}_3^{alt} = {}^{\hat{P}(M|C,d^*)}/_{\hat{P}(M|C,d)\hat{P}(d|C)}$ substituted for $\hat{w}_3 = {}^{\hat{P}(d^*|C,M)}/_{\hat{P}(d|C,M)\hat{P}(d^*|C)}$. Point estimates constructed using $\hat{w}_3^{alt}$ may differ from those based on $\hat{w}_3$ due to sampling error or misspecification bias, but with full population data and correct models for all the terms that compose the weights, these two different implementations would yield exactly the same results–that is, they are asymptotically equivalent if there is no model misspecification.

Appendix G

Nonparametric Identification of CDEs with Exposure-induced Confounding

If assumptions (d.i) to (d.iv) are all satisfied, the $CDE\left(d, d^{*}, m\right)$ can be nonparametrically identified as follows:

$$
\begin{aligned}
CDE\left(d, d^{*}, m\right) &= \mathbb{E}\left[Y\left(d, m\right) - Y\left(d^{*}, m\right)\right] \\
&= \mathbb{E}\left[Y\left(d, m\right)\right] - \mathbb{E}\left[Y\left(d^{*}, m\right)\right] && \textit{by properties of } \mathbb{E}\left[\cdot\right] \\
&= \mathbb{E}_{C}\left[\mathbb{E}\left[Y\left(d, m\right)|C\right] - \mathbb{E}\left[Y\left(d^{*}, m\right)|C\right]\right] && \textit{by iterated expectations} \\
&= \mathbb{E}_{C}\left[\mathbb{E}\left[Y\left(d, m\right)|C, d\right] - \mathbb{E}\left[Y\left(d^{*}, m\right)|C, d^{*}\right]\right] && \textit{by (d.i) and (d.iii)} \\
&= \mathbb{E}_{C}\left[\mathbb{E}_{L|C,d}\left[\mathbb{E}\left[Y\left(d, m\right)|C, d, L\right]\right]\right] \\
&\qquad - \mathbb{E}_{C}\left[\mathbb{E}_{L|C,d^{*}}\left[\mathbb{E}\left[Y\left(d^{*}, m\right)|C, d^{*}, L\right]\right]\right] && \textit{by iterated expectations} \\
&= \mathbb{E}_{C}\left[\mathbb{E}_{L|C,d}\left[\mathbb{E}\left[Y\left(d, m\right)|C, d, L, m\right]\right]\right] \\
&\qquad - \mathbb{E}_{C}\left[\mathbb{E}_{L|C,d^{*}}\left[\mathbb{E}\left[Y\left(d^{*}, m\right)|C, d^{*}, L, m\right]\right]\right] && \textit{by (d.ii) and (d.iii)} \\
&= \mathbb{E}_{C}\left[\mathbb{E}_{L|C,d}\left[\mathbb{E}\left[Y|C, d, L, m\right]\right] - \mathbb{E}_{L|C,d^{*}}\left[\mathbb{E}\left[Y|C, d^{*}, L, m\right]\right]\right]. && \textit{by (d.iv)}
\end{aligned}
$$

The final equality in this derivation gives an expression for the $CDE(d, d^{*}, m)$ that is defined solely in terms of the observable data $\{Y, C, D, L, M\}$, where L is an exposure-induced confounder. We arrived at this expression by starting with the formal definition of the controlled direct effect and then by invoking assumptions (d.i) to (d.iv) together with the same properties of expected values that we outlined previously in Appendix A. This expression can also be written as follows:

$$
\sum_{l,c}\left(\mathbb{E}\left[Y|c, d, l, m\right] P\left(l|c, d\right) - \mathbb{E}\left[Y|c, d^{*}, l, m\right] P\left(l|c, d^{*}\right)\right) P\left(c\right),
$$

because $\mathbb{E}_{C}\left[\cdot\right]$ is just an expected value taken over the distribution C and $\mathbb{E}_{L|C,D}\left[\cdot\right]$ is an expected value taken over the conditional distribution of L given C and D . If C or L were continuous, the probability-weighted sums would be replaced with density-weighted integrals.

Appendix H

Nonparametric Identification of the Interventional Direct Effect

Under assumptions (e.i) to (e.v), the $IDE\,(d, d^*)$ can be nonparametrically identified as follows:

$$\begin{aligned}
IDE\,(d, d^*) &= \mathbb{E}\left[Y\left(d, \mathcal{M}\left(d^*|C\right)\right) - Y\left(d^*, \mathcal{M}\left(d^*|C\right)\right)\right] \\
&= \sum_{m,c} \mathbb{E}\left[Y\left(d, m\right) - Y\left(d^*, m\right)|c\right] P\left(M\left(d^*\right) = m|c\right) P\left(c\right) && \textit{by iterated exp.} \\
&= \sum_{m,c} \left(\mathbb{E}\left[Y\left(d, m\right)|c\right] - \mathbb{E}\left[Y\left(d^*, m\right)|c\right]\right) P\left(M\left(d^*\right) = m|c\right) P\left(c\right) && \textit{by prop. of } \mathbb{E}\left[\cdot\right] \\
&= \sum_{m,c} \left(\mathbb{E}\left[Y\left(d, m\right)|c, d\right] - \mathbb{E}\left[Y\left(d^*, m\right)|c, d^*\right]\right) \\
&\quad \times P\left(M\left(d^*\right) = m|c, d^*\right) P\left(c\right) && \textit{by (e.i), (e.iii), (e.iv)} \\
&= \sum_{m,c} \mathbb{E}\left[Y\left(d, m\right)|c, d, L\right] P\left(M\left(d^*\right) = m|c, d^*\right) P\left(c\right) \\
&\quad - \sum_{m,c} \mathbb{E}\left[Y\left(d^*, m\right)|c, d^*, L\right] P\left(M\left(d^*\right) = m|c, d^*\right) P\left(c\right) && \textit{by distributive law} \\
&= \sum_{m,c} \mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y\left(d, m\right)|c, d, L\right]\right] P\left(M\left(d^*\right) = m|c, d^*\right) P\left(c\right) \\
&\quad - \sum_{m,c} \mathbb{E}_{L|c,d^*}\left[\mathbb{E}\left[Y\left(d^*, m\right)|c, d^*, L\right]\right] P\left(M\left(d^*\right) = m|c, d^*\right) P\left(c\right) && \textit{by iterated exp.} \\
&= \sum_{m,c} \mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y\left(d, m\right)|c, d, L, m\right]\right] P\left(M\left(d^*\right) = m|c, d^*\right) P\left(c\right) \\
&\quad - \sum_{m,c} \mathbb{E}_{L|c,d^*}\left[\mathbb{E}\left[Y\left(d^*, m\right)|c, d^*, L, m\right]\right] P\left(M\left(d^*\right) = m|c, d^*\right) P\left(c\right) && \textit{by (e.ii), (e.iv)} \\
&= \sum_{m,c} \mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y|c, d, L, m\right]\right] P\left(M = m|c, d^*\right) P\left(c\right) \\
&\quad - \sum_{m,c} \mathbb{E}_{L|c,d^*}\left[\mathbb{E}\left[Y|c, d^*, L, m\right]\right] P\left(M = m|c, d^*\right) P\left(c\right) && \textit{by (c.v)}
\end{aligned}$$

The final equality in this derivation gives an expression for the $IDE\,(d, d^*)$ that is defined only in terms of the observable data $\{Y, C, D, L, M\}$ and does not involve any functional form restrictions. We arrived at this expression by starting with the formal definition of the interventional direct effect and then by invoking assumptions (e.i) to (e.v) together with the same properties of expected values that we previously outlined in Appendix A. This expression can also be written as follows:

$$IDE\left(d, d^{*}\right) = \sum_{l,m,c} \left(\mathbb{E}\left[Y|c, d, l, m\right] P\left(l|c, d\right) - \mathbb{E}\left[Y|c, d^{*}, l, m\right] P\left(l|c, d^{*}\right)\right) P\left(m|c, d^{*}\right) P\left(c\right)$$

because of the distributive law and because $\mathbb{E}_{L|C,D}\left[\cdot\right]$ is an expected value taken over the conditional distribution of the exposure-induced confounder L, given the baseline confounders C and exposure D. The probability-weighted sums in this expression would be replaced with density-weighted integrals if C, L, or M were continuous.

Finally, note that when L is an empty set, that is, when there are no exposure-induced confounders, the nonparametric identification formula for the interventional direct effect can be simplified as follows:

$$\begin{aligned} IDE\left(d, d^{*}\right) &= \sum_{l,m,c} \left(\mathbb{E}\left[Y|c, d, l, m\right] P\left(l|c, d\right) - \mathbb{E}\left[Y|c, d^{*}, l, m\right] P\left(l|c, d^{*}\right)\right) P\left(m|c, d^{*}\right) P\left(c\right) \\ &= \sum_{m,c} \left(\mathbb{E}\left[Y|c, d, m\right] - \mathbb{E}\left[Y|c, d^{*}, m\right]\right) P\left(m|c, d^{*}\right) P\left(c\right), \end{aligned}$$

where the final expression is equal to the nonparametric identification formula for the natural direct effect. Thus, in the absence of exposure-induced confounding, interventional and natural direct effects are given by the same function of observable data, provided their other identification assumptions are met.

Appendix I

Nonparametric Identification of the Interventional Indirect Effect

Under assumptions (e.i) to (e.v), the $IIE\left(d,d^{*}\right)$ can be nonparametrically identified as follows:

$$\begin{aligned}
NIE\left(d,d^{*}\right) &= \mathbb{E}\left[Y\left(d,\mathcal{M}\left(d|C\right)\right)-Y\left(d,\mathcal{M}\left(d^{*}|C\right)\right)\right] \\
&= \sum_{m,c}\mathbb{E}\left[Y\left(d,m\right)|c\right]P\left(M\left(d\right)=m|c\right)P\left(c\right) \\
&\qquad -\sum_{m,c}\mathbb{E}\left[Y\left(d,m\right)|c\right]P\left(M\left(d^{*}\right)=m|c\right)P\left(c\right) && \textit{by iterated exp.} \\
&= \sum_{m,c}\mathbb{E}\left[Y\left(d,m\right)|c,d\right]P\left(M\left(d\right)=m|c,d\right)P\left(c\right) \\
&\qquad -\sum_{m,c}\mathbb{E}\left[Y\left(d,m\right)|c,d\right]P\left(M\left(d^{*}\right)=m|c,d^{*}\right)P\left(c\right) && \textit{by (e.i), (e.iii), (e.iv)} \\
&= \sum_{m,c}\mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y\left(d,m\right)|c,d,L\right]\right]P\left(M\left(d\right)=m|c,d\right)P\left(c\right) \\
&\qquad -\sum_{m,c}\mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y\left(d,m\right)|c,d,L\right]\right]P\left(M\left(d^{*}\right)=m|c,d^{*}\right)P\left(c\right) && \textit{by iterated exp.} \\
&= \sum_{m,c}\mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y\left(d,m\right)|c,d,L,m\right]\right]P\left(M\left(d\right)=m|c,d\right)P\left(c\right) \\
&\qquad -\sum_{m,c}\mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y\left(d,m\right)|c,d,L,m\right]\right]P\left(M\left(d^{*}\right)=m|c,d^{*}\right)P\left(c\right) && \textit{by (c.ii), (c.iv)} \\
&= \sum_{m,c}\mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y|c,d,L,m\right]\right]P\left(M=m|c,d\right)P\left(c\right) \\
&\qquad -\sum_{m,c}\mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y|c,d,L,m\right]\right]P\left(M=m|c,d^{*}\right)P\left(c\right) && \textit{by (c.v)} \\
&= \sum_{m,c}\left(P\left(m|c,d\right)-P\left(m|c,d^{*}\right)\right)\mathbb{E}_{L|c,d}\left[\mathbb{E}\left[Y|c,d,L,m\right]\right]P\left(c\right) && \textit{by distributive law}
\end{aligned}$$

where the final equality in this derivation gives an expression for the $IIE\left(d,d^{*}\right)$ that is defined only in terms of the observable data $\{Y,C,D,L,M\}$ and without any functional form restrictions. We arrived at this expression by starting with the formal definition of the interventional indirect effect in terms of counterfactuals and then by invoking assumptions (e.i) to (e.v). We also relied on several properties of expected values, which were previously described in Appendix A. This expression can also be written as follows:

$$IIE\left(d,d^{*}\right)=\sum_{l,m,c}\left(P\left(m|c,d\right)-P\left(m|c,d^{*}\right)\right)\mathbb{E}\left[Y|c,d,l,m\right]P\left(l|c,d\right)P\left(c\right),$$

because $\mathbb{E}_{L|C,D}\left[\cdot\right]$ is an expected value taken over the conditional distribution of L, given C and D. If C, L or M were continuous, the probability-weighted sums would be replaced with density-weighted integrals.

When there are no exposure-induced confounders and thus L is an empty set, the nonparametric identification formula for the interventional indirect effect can be simplified as follows:

$$\begin{aligned}IIE\left(d,d^{*}\right)&=\sum_{l,m,c}\left(P\left(m|c,d\right)-P\left(m|c,d^{*}\right)\right)\mathbb{E}\left[Y|c,d,l,m\right]P\left(l|c,d\right)P\left(c\right)\\&=\sum_{l,m,c}\left(P\left(m|c,d\right)-P\left(m|c,d^{*}\right)\right)\mathbb{E}\left[Y|c,d,m\right]P\left(c\right),\end{aligned}$$

where the final expression is equivalent to the nonparametric identification formula for the natural indirect effect. Interventional and natural indirect effects are therefore given by the same function of observable data in the absence of exposure-induced confounding, provided their other identification assumptions are met.

Appendix J

Bias Formulas for PSEs with $K \geq 2$ Causally Ordered Mediators

Suppose that the causal effect of the exposure on the outcome operates through K causally ordered mediators $M_1, M_2, \ldots M_K$, and that there exists an unobserved confounder U that affects the exposure D, the outcome Y, and the mediators $\{M_1, M_2, \ldots M_K\}$. Let $\mathbf{M}_0 = \varnothing$ denote an empty set, let $\mathbf{M}_k = \{M_1, M_2, \ldots, M_k\}$ denote a cumulative set of mediators up through M_k, and let $\mathbf{M}_k(d) = \{M_1(d), M_2(d), \ldots M_k(d)\}$ represent their potential values under exposure d, where $M_k(d) = M_k(d, M_1(d), M_2(d, M_1(d)), \ldots)$ by definition. With this notation, the PSEs in Equation 5.10 can all be expressed as functions of natural direct and indirect effects, where

$$\begin{aligned}
NDE_{\mathbf{M}_k}(d, d^*) &= PSE_{D \,\to\, Y}(d, d^*) + \sum_{l=k+1}^{K} PSE_{D \,\to\, M_l \,\rightsquigarrow\, Y}(d, d^*) \\
NIE_{\mathbf{M}_k}(d, d^*) &= \sum_{l=1}^{k} PSE_{D \,\to\, M_l \,\rightsquigarrow\, Y}(d, d^*),
\end{aligned}$$

and the PSEs of interest can be expressed as follows:

$$\begin{aligned}
PSE_{D \,\to\, Y}(d, d^*) &= NDE_{\mathbf{M}_k}(d, d^*) \\
PSE_{D \,\to\, M_k \,\rightsquigarrow\, Y}(d, d^*) &= NIE_{\mathbf{M}_k}(d, d^*) - NIE_{\mathbf{M}_{k-1}}(d, d^*).
\end{aligned}$$

Estimators for the PSEs that do not adjust for U are therefore subject to the following biases:

$$\text{Bias}\left(\widehat{PSE}_{D \to Y}(d, d^*)\right) = \text{Bias}\left(\widehat{NDE}_{\mathbf{M}_k}(d, d^*)\right) \tag{J.1}$$

$$\text{Bias}\left(\widehat{PSE}_{D \to M_k \to Y}(d, d^*)\right) = \text{Bias}\left(\widehat{NIE}_{\mathbf{M}_k}(d, d^*)\right) - \text{Bias}\left(\widehat{NIE}_{\mathbf{M}_{k-1}}(d, d^*)\right), \tag{J.2}$$

where the bias formulas for the natural effects on the right-hand side are given in Section 3.7.

Now suppose that the unobserved variable U affects only the exposure and the outcome, but not the mediators directly, and that U is binary, that $\mathbb{E}[Y|c, d, m_1, m_2, \ldots, m_k, U = 1] - \mathbb{E}[Y|c, d, m_1, m_2, \ldots, m_k, U = 0]$ is constant across levels of C, D, and $\mathbf{M}_k$ for each $k \in \{1, 2, \ldots, K\}$, and that $P(U = 1|c, d) - P(U = 1|c, d^*)$ is constant across levels of C. Under these assumptions, the bias formulas J.1 and J.2 reduce to the following expressions:

$$\begin{aligned}\text{Bias}\left(\widehat{PSE}_{D\to Y}\left(d,d^*\right)\right) &= \delta_{UY|C,D,\mathbf{M}_k}\times\delta_{DU|C}\\ \text{Bias}\left(\widehat{PSE}_{D\to M_k\to Y}\left(d,d^*\right)\right) &= 0,\quad k\in\{1,2,\ldots,K\},\end{aligned}$$

where $\delta_{UY|C,D,\mathbf{M}_k}$ and $\delta_{DU|C}$ are defined as

$$\begin{aligned}\delta_{UY|C,D,\mathbf{M}_k} &= \mathbb{E}\left[Y|c,d,m_1,m_2,\ldots,m_K,U=1\right]-\mathbb{E}\left[Y|c,d,m_1,m_2,\ldots,m_K,U=0\right]\\ \delta_{DU|C} &= P\left(U=1|c,d\right)-P\left(U=1|c,d^*\right).\end{aligned}$$

Next, suppose that the unobserved variable U affects only the mediators and the outcome, but not the exposure, and that U is binary, that $\mathbb{E}\left[Y|c,d,m_1,m_2,\ldots,m_k,U=1\right]-\mathbb{E}\left[Y|c,d,m_1,m_2,\ldots,m_k,U=0\right]$ is constant across levels of C, D, and $\mathbf{M}_k$ for each $k\in\{1,2,\ldots K\}$, and that $P\left(U=1|c,d,m_1,m_2,\ldots,m_k\right)-P\left(U=1|c,d^*,m_1,m_2,\ldots,m_k\right)$ is constant across levels of C and $\mathbf{M}_k$ for each $k\in\{1,2,\ldots K\}$. Under these assumptions, the bias formulas J.1 and J.2 reduce to the following expressions:

$$\begin{aligned}\text{Bias}\left(\widehat{PSE}_{D\to Y}\left(d,d^*\right)\right) &= \delta_{UY|C,D,\mathbf{M}_k}\times\delta_{DU|C,\mathbf{M}_k}\\ \text{Bias}\left(\widehat{PSE}_{D\to M_1\to Y}\left(d,d^*\right)\right) &= -\delta_{UY|C,D,M_1}\times\delta_{DU|C,D,M_1}\\ \text{Bias}\left(\widehat{PSE}_{D\to M_k\to Y}\left(d,d^*\right)\right) &= \delta_{UY|C,D,\mathbf{M}_{k-1}}\times\delta_{DU|C,\mathbf{M}_{k-1}}-\delta_{UY|C,D,\mathbf{M}_k}\times\delta_{DU|C,\mathbf{M}_k},\quad k\in\{2,\ldots K\},\end{aligned}$$

where $\delta_{UY|C,D,\mathbf{M}_k}$ is defined as before, and $\delta_{DU|C,\mathbf{M}_k}$ is defined as

$$\delta_{DU|C,\mathbf{M}_k}=P\left(U=1|c,d,m_1,\ldots,m_k\right)-P\left(U=1|c,d^*,m_1,\ldots,m_k\right).$$

Finally, suppose that the unobserved variable U affects only the exposure and the mediators, but not the outcome directly, and that U is binary, that $P\left(U=1|c,d\right)-P\left(U=1|c,d^*\right)$ is constant across levels of C, that $P\left(m_1,\ldots,m_k|c,d,U=1\right)-P\left(m_1,\ldots,m_k|c,d,U=0\right)$ is constant across levels of D for each $k\in\{1,2,\ldots K\}$, and that the controlled direct effect of the exposure on the outcome with respect to $\mathbf{M}_k$ does not depend on $(m_1,m_2\ldots,m_K)$ within levels of the baseline confounders. Under these assumptions, the bias formulas J.1 and J.2 reduce to the following expressions:

$$\begin{aligned}\text{Bias}\left(\widehat{PSE}_{D\to Y}\left(d,d^*\right)\right) &= 0,\\ \text{Bias}\left(\widehat{PSE}_{D\to M_1\to Y}\left(d,d^*\right)\right) &= \delta_{DU|C}\times\delta_{UM_1Y|C},\\ \text{Bias}\left(\widehat{PSE}_{D\to M_k\to Y}\left(d,d^*\right)\right) &= \delta_{DU|C}\times\left(\delta_{U\mathbf{M}_kY|C}-\delta_{U\mathbf{M}_{k-1}Y|C}\right),\quad k\in\{2,\ldots K\},\end{aligned}$$

where $\delta_{DU|C}$ is defined as before, and $\delta_{U\mathbf{M}_kY|C}$ is defined as

$$\delta_{U\mathbf{M}_kY|C}=\sum_{m_1,m_2,\ldots,m_k,c}\mathbb{E}\left[Y|c,d,m_1,\ldots,m_k\right]\left(P\left(m_1,\ldots,m_k|c,d,U=1\right)-P\left(m_1,\ldots,m_k|c,d,U=0\right)\right)P\left(c\right).$$

Bibliography

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72(1):1–19.

Acharya, A., Blackwell, M., and Sen, M. (2016). Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects. *American Political Science Review*, 110(3):512–529.

Acharya, A., Blackwell, M., and Sen, M. (2018). Analyzing Causal Mechanisms in Survey Experiments. *Political Analysis*, 26(4):357–378.

Adams, P., Hurd, M. D., McFadden, D., Merrill, A., and Ribeiro, T. (2003). Healthy, Wealthy, and Wise? Tests for Direct Causal Paths between Health and Socioeconomic Status. *Journal of Econometrics*, 112(1):3–56.

Agresti, A. (2012). *Categorical Data Analysis*. John Wiley & Sons, New York, NY.

Aldrich, J. H. and Nelson, F. D. (1984). *Linear Probability, Logit, and Probit Models*. SAGE Publications, Thousand Oaks, CA.

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the Origins of Gender Roles: Women and the Plough. *The Quarterly Journal of Economics*, 128(2):469–530.

Allison, P. D. (2009). *Fixed Effects Regression Models*. SAGE Publications, Thousand Oaks, CA.

Almirall, D., Ten Have, T., and Murphy, S. A. (2010). Structural Nested Mean Models for Assessing Time-Varying Effect Moderation. *Biometrics*, 66(1):131–139.

Alwin, D. F. and Hauser, R. M. (1975). The Decomposition of Effects in Path Analysis. *American Sociological Review*, 40(1):37–47.

Andrews, R. M. and Didelez, V. (2021). Insights into the Cross-world Independence Assumption of Causal Mediation Analysis. *Epidemiology*, 32(2):209–219.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455.

Angrist, J. D. and Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, 15(4):69–85.

Angrist, J. D. and Pischke, J. S. (2009). *Mostly Harmless Econometrics*. Princeton University Press, Princeton, NJ.

Angrist, J. D. and Pischke, J. S. (2014). *Mastering'metrics: The Path from Cause to Effect*. Princeton University Press, Princeton, NJ.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association*, 116(536):1716–1730.

Athey, S. and Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, 11(1):685–725.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of Path-Specific Effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 357–363, San Francisco, CA. Morgan Kaufmann Publishers.

Avison, W. R., McLeod, J. D., and Pescosolido, B. A. (2007). *Mental Health, Social Mirror*. Springer Publishing, New York, NY.

Balgi, S., Daoud, A., Pena, J. M., Wodtke, G. T., and Zhou, J. (2024). Deep Learning With DAGs. *arXiv preprint arXiv:2401.06864*.

Barnow, B. S. (1987). The Impact of CETA Programs on Earnings: A Review of the Literature. *Journal of Human Resources*, 22:157–193.

Baron, R. M. and Kenny, D. A. (1986). The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.

Becker, G. S. (2010). *The Economics of Discrimination*. University of Chicago Press, Chicago, IL.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on Treatment Effects After Selection Among High-Dimensional Controls. *Review of Economic Studies*, 81(2):608–650.

Benkeser, D. and Ran, J. (2021). Nonparametric Inference for Interventional Effects with Multiple Mediators. *Journal of Causal Inference*, 9(1):172–189.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous Analysis of Lasso and Dantzig Selector. *The Annals of Statistics*, 37(4):1705–1732.

Blackwell, M., Glynn, A., and Hilbig, H. (2024). Estimating Controlled Direct Effects with Panel Data: An Application to Reducing Support for Discriminatory Policies.

Blackwell, M., Glynn, A., Hilbig, H., and Phillips, C. H. (2022). Difference-in-differences Designs for Controlled Direct Effects.

Blackwell, M. and Yamauchi, S. (2021). Adjusting for Unmeasured Confounding in Marginal Structural Models with Propensity-score Fixed Effects. *arXiv preprint arXiv:2105.03478*.

Blalock, H. M. (1971). *Causal Models in the Social Sciences.* Aldine-Atherton, Chicago, IL.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. (1997). The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study. *Journal of Human Resources*, 32(3):549–576.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* John Wiley & Sons, New York, NY.

Bonate, P. L. (2000). *Analysis of Pretest-Posttest Designs.* Chapman and Hall/CRC, New York, NY.

Boserup, E. (1970). *Woman's Role in Economic Development.* George Allen and Unwin, London.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90(430):443–450.

Brader, T., Valentino, N. A., and Suhay, E. (2008). What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat. *American Journal of Political Science*, 52(4):959–978.

Brady, D. (2009). *Rich Democracies, Poor People: How Politics Explain Poverty.* Oxford University Press, Oxford.

Brand, J. E., Zhou, X., and Xie, Y. (2023). Recent Developments in Causal Inference and Machine Learning. *Annual Review of Sociology*, 49(1):81–110.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.

Bullock, J. G. and Ha, S. E. (2011). Mediation Analysis Is Harder Than It Looks. In Druckman, J. N., Greene, D. P., Kuklinski, J. H., and Lupia, A., editors, *Cambridge Handbook of Experimental Political Science*, pages 508–522. Cambridge University Press, Cambridge.

Bureau of Labor Statistics, U. S. (2019). National Longitudinal Survey of Youth 1979 Cohort, 1979-2016 (Rounds 1-27). Technical report, US Department of Labor.

Burgess, S., Daniel, R. M., Butterworth, A. S., Thompson, S. G., and Consortium, E.-I. (2015). Network Mendelian Randomization: Using Genetic Variants as Instrumental Variables to Investigate Mediation in Causal Pathways. *International Journal of Epidemiology*, 44(2):484–495.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, 225(2):200–230.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3):414–427.

Campbell, D. T. and Stanley, J. C. (2015). *Experimental and Quasi-Experimental Designs for Research.* Ravenio Books.

Card, D. (1999). The Causal Effect of Education on Earnings. volume 3 of *Handbook of Labor Economics*, pages 1801–1863. Elsevier, Amsterdam.

Card, D. (2001). Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, 69(5):1127–1160.

Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review*, 84(4):772–793.

Chen, X. and White, H. (1999). Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators. *IEEE Transactions on Information Theory*, 45(2):682–691.

Cheng, C., Spiegelman, D., and Li, F. (2023). Mediation Analysis in the Presence of Continuous Exposure Measurement Error. *Statistics in Medicine*, 42(11):1669–1686.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21(1):C1–C68.

Cole, S. R. and Hernan, M. A. (2008). Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology*, 168(6):656–664.

Cook, T. D. (2002). Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them. *Educational Evaluation and Policy Analysis*, 24(3):175–199.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. (2015). Causal Mediation Analysis with Multiple Mediators. *Biometrics*, 71(1):1–14.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

de Chaisemartin, C. and D'Haultfoeuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–2996.

De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., and Micali, N. (2015). Mediation Analysis With Intermediate Confounding: Structural Equation Modeling Viewed Through the Causal Inference Lens. *American Journal of Epidemiology*, 181(1):64–80.

Diaz, I., Hejazi, N. S., Rudolph, K. E., and van der Laan, M. J. (2021). Nonparametric Efficient Causal Mediation with Intermediate Confounders. *Biometrika*, 108(3):627–641.

Diaz, I. and van der Laan, M. J. (2013). Targeted Data Adaptive Estimation of the Causal Dose Response Curve. *Journal of Causal Inference*, 1(2):171–192.

Didelez, V., Dawid, P., and Geneletti, S. (2012). Direct and Indirect Effects of Sequential Treatments. *arXiv preprint arXiv:1206.6840*.

Dippel, C., Gold, R., Heblich, S., and Pinto, R. (2017). Instrumental Variables and Causal Mechanisms: Unpacking the Effect of Trade on Workers and Voters. Working Paper 23209, National Bureau of Economic Research.

Dukes, O., Shpitser, I., and Tchetgen Tchetgen, E. J. (2023). Proximal Mediation Analysis. *Biometrika*, 110(4):973–987.

Duncan, O. D. (1966). Path Analysis: Sociological Examples. *American Journal of Sociology*, 72(1):1–16.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap.* Chapman and Hall/CRC, New York, NY.

Elwert, F. (2013). Graphical Causal Models. In Morgan, S. L., editor, *Handbook of Causal Analysis for Social Research*, Handbooks of Sociology and Social Research, pages 245–273. Springer, Dordrecht.

Elwert, F. and Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, 40:31–53.

Farbmacher, H., Huber, M., Laffars, L., Langen, H., and Spindler, M. (2022). Causal Mediation Analysis with Double Machine Learning. *The Econometrics Journal*, 25(2):277–300.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models.* SAGE Publications, Thousand Oaks, CA.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of statistics*, 29:1189–1232.

Frolich, M. and Huber, M. (2017). Direct and Indirect Treatment Effects-Causal Chains and Mediation Analysis with Instrumental Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(5):1645–1666.

Gaddis, S. M. (2015). Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market. *Social Forces*, 93(4):1451–1479.

Gaddis, S. M. (2018). *An Introduction to Audit Studies in the Social Sciences.* Springer Publishing, New York, NY.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, Cambridge.

Geneletti, S. (2007). Identifying Direct and Indirect Effects in a Non-counterfactual Framework. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):199–215.

Glazer, N. and Moynihan, D. P. (1970). *Beyond the Melting Pot: The Negroes, Puerto Ricans, Jews, Italians, and Irish of New York City.* MIT Press, Cambridge, MA.

Glynn, A. N. (2021). Advances in Experimental Mediation Analysis. In Druckman, J. N. and Green, D. P., editors, *Advances in Experimental Political Science*, pages 257–270. Cambridge University Press, Cambridge.

Goldberger, A. S. (1972). Structural Equation Methods in the Social Sciences. *Econometrica*, 40(6):979–1001.

Goldberger, A. S. and Duncan, O. D. (1973). *Structural Equation Models in the Social Sciences.* Seminar Press, New York, NY.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning.* MIT Press, Cambridge, MA.

Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometrics*, 225(2):254–277.

Green, D. P., Ha, S. E., and Bullock, J. G. (2010). Enough Already About "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose. *The ANNALS of the American Academy of Political and Social Science*, 628(1):200–208.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, 24:395–419.

Halaby, C. N. (2004). Panel Models in Sociological Research: Theory into Practice. *Annual Review of Sociology*, 30:507–544.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer Publishing, New York, NY.

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach.* Methodology in the Social Sciences. Guilford Publications, New York, NY.

Heaton, T. B. (1991). Time-Related Determinants of Marital Dissolution. *Journal of Marriage and Family*, 53(2):285–295.

Heckman, J. J. (1998). Detecting Discrimination. *Journal of Economic Perspectives*, 12(2):101–116.

Heckman, J. J., Humphries, J. E., and Veramendi, G. (2018). The Nonmarket Benefits of Education and Ability. *Journal of Human Capital*, 12(2):282–304.

Heckman, J. J. and Smith, J. A. (1995). Assessing the Case for Social Experiments. *Journal of Economic Perspectives*, 9(2):85–110.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference: What If.* Chapman & Hall/CRC, Boca Raton, FL.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hines, O., Dukes, O., Diaz-Ordaz, K., and Vansteelandt, S. (2022). Demystifying Statistical Learning Based on Efficient Influence Functions. *The American Statistician*, 76(3):292–304.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Holm, A. and Breen, R. (2023). Causal Mediation in Panel Data: Estimation Based on Difference in Differences. SocArXiv kwscz, Center for Open Science.

Holzer, H. J. (2013). Workforce Development Programs. In Bailey, M. J. and Danziger, S., editors, *Legacies of the War on Poverty*, National Poverty Center Series on Poverty and Public Policy, pages 121–150. Russell Sage Foundation, New York, NY.

Hong, G. (2015). *Causality in a Social World: Moderation, Mediation and Spill-over.* John Wiley & Sons, New York, NY.

Hong, G., Deutsch, J., and Hill, H. D. (2015). Ratio-of-Mediator-Probability Weighting for Causal Mediation Analysis in the Presence of Treatment-by-Mediator Interaction. *Journal of Educational and Behavioral Statistics*, 40(3):307–340.

Hout, M. (2012). Social and Economic Returns to College Education in the United States. *Annual Review of Sociology*, 38:379–400.

Huber, M., Schelker, M., and Strittmatter, A. (2022). Direct and Indirect Effects based on Changes-in-Changes. *Journal of Business & Economic Statistics*, 40(1):432–443.

Imai, K., Keele, L., and Tingley, D. (2010a). A General Approach to Causal Mediation Analysis. *Psychological Methods*, 15(4):309.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review*, 105(4):765–789.

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

Imai, K. and Kim, I. S. (2021). On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. *Political Analysis*, 29(3):405–415.

Imai, K., Tingley, D., and Yamamoto, T. (2013). Experimental Designs for Identifying Causal Mechanisms. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 176(1):5–51.

Imai, K. and Yamamoto, T. (2013). Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments. *Political Analysis*, 21(2):141–171.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: an Introduction*. Cambridge University Press.

Jahoda, M. (1982). *Employment and Unemployment: A Social-Psychological Analysis*. Cambridge University Press, Cambridge.

Javaloy, A., Sanchez-Martin, P., and Valera, I. (2023). Causal Normalizing Flows: From Theory to Practice. In *Advances in Neural Information Processing Systems*, volume 36, pages 58833–58864. Curran Associates, Inc.

Jiang, Z. and VanderWeele, T. (2019). Causal Mediation Analysis in the Presence of a Misclassified Binary Exposure. *Epidemiologic Methods*, 8(1):1–12.

Jiang, Z. and VanderWeele, T. J. (2015). Causal Mediation Analysis in the Presence of a Mismeasured Outcome. *Epidemiology*, 26(1):e8–e9.

Jun, S. J., Pinkse, J., Xu, H., and Yildiz, N. (2016). Multiple Discrete Endogenous Variables in Weakly–Separable Triangular Models. *Econometrics*, 4(1):7.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523–539.

Karlson, K. B., Holm, A., and Breen, R. (2012). Comparing Regression Coefficients Between Same-sample Nested Models Using Logit and Probit: A New Method. *Sociological Methodology*, 42(1):286–313.

Kennedy, E. H. (2016). Semiparametric Theory and Empirical Processes in Causal Inference. In He, H., Wu, P., and Chen, D.-G., editors, *Statistical Causal Inferences and Their Applications in Public Health Research*, ICSA Book Series in Statistics, pages 141–167. Springer Publishing, New York, NY.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-Parametric Methods for Doubly Robust Estimation of Continuous Treatment Effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245.

Kline, P., Rose, E. K., and Walters, C. R. (2022). Systemic Discrimination Among Large U.S. Employers. *The Quarterly Journal of Economics*, 137(4):1963–2036.

Krueger, A. O. (1963). The Economics of Discrimination. *Journal of Political Economy*, 71(5):481–486.

Lange, T. and Hansen, J. V. (2011). Direct and Indirect Effects in a Survival Context. *Epidemiology*, 22(4):575–581.

le Cessie, S., Debeij, J., Rosendaal, F. R., Cannegieter, S. C., and Vandenbrouckea, J. P. (2012). Quantification of Bias in Direct Effects Estimates Due to Different Types of Measurement Error in the Mediator. *Epidemiology*, 23(4):551–560.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving Propensity Score Weighting Using Machine Learning. *Statistics in Medicine*, 29(3):337–346.

Lee, J. (2011). Pathways from Education to Depression. *Journal of Cross-Cultural Gerontology*, 26(2):121–135.

Lin, Q., Nutall, A. K., Zhang, Q., and Frank, K. A. (2023). How Do Unobserved Confounding Mediators and Measurement Error Impact Estimated Mediation Effects and Corresponding Statistical Inferences? Introducing the R Package ConMed for Sensitivity Analysis. *Psychological Methods*, 28(2):339–358.

Lin, S.-H. and VanderWeele, T. (2017). Interventional Approach for Path-Specific Effects. *Journal of Causal Inference*, 5(1):20150027.

Lumley, T. (2011). *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons, New York, NY.

Lundberg, I., Johnson, R., and Stewart, B. M. (2021). What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, 86(3):532–565.

Lupu, N. and Peisakhin, L. (2017). The Legacy of Political Violence across Generations. *American Journal of Political Science*, 61(4):836–851.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. Routledge, New York, NY.

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation Analysis. *Annual Review of Psychology*, 58:593–614.

Maimon, O. Z. and Rokach, L. (2014). *Data Mining with Decision Trees: Theory and Applications*, volume 81 of *Series in Machine Perception And Artificial Intelligence*. World Scientific, Singapore.

Marcus, G. E., Neuman, W. R., and MacKuen, M. (2000). *Affective Intelligence and Political Judgment*. University of Chicago Press, Chicago, IL.

Marcus, G. E., Sullivan, J. L., Theiss-Morse, E., and Stevens, D. (2005). The Emotional Foundations of Political Cognition: The Impact of Extrinsic Anxiety on the Formation of Political Tolerance Judgments. *Political Psychology*, 26(6):949–963.

Marshall, M. and Jaggers, K. (2007). Polity IV Project: Dataset Users' Manual.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, 9(4):403–425.

McCullagh, P. (1989). *Generalized Linear Models*. Routledge, New York, NY.

Meyer, B. D. (1995). Natural and Quasi-Experiments in Economics. *Journal of Business & Economic Statistics*, 13(2):151–161.

Miech, R. A., Caspi, A., Moffitt, T. E., Wright, B. R. E., and Silva, P. A. (1999). Low Socioeconomic Status and Mental Disorders: A Longitudinal Study of Selection and Causation during Young Adulthood. *American Journal of Sociology*, 104(4):1096–1131.

Miles, C. H., Shpitser, I., Kanki, P., Meloni, S., and Tchetgen Tchetgen, E. J. (2020). On Semiparametric Estimation of a Path-Specific Effect in the Presence of Mediator-Outcome Confounding. *Biometrika*, 107(1):159–172.

Miller, J. M. (2007). Examining the Mediators of Agenda Setting: A New Experimental Paradigm Reveals the Role of Emotions. *Political Psychology*, 28(6):689–717.

Mills, C. and D'Mello, S. (2014). On the Validity of the Autobiographical Emotional Memory Task for Emotion Induction. *PloS one*, 9(4):e95837.

Mirowsky, J. (2003). *Education, Social Status, and Health*. Routledge, New York, NY.

Mirowsky, J. and Ross, C. E. (1990). Control or Defense? Depression and the Sense of Control over Good and Bad Outcomes. *Journal of Health and Social Behavior*, 31(1):71–86.

Montgomery, D. C. (2019). *Design and Analysis of Experiments, 10th Edition*. John Wiley & Sons, New York, NY.

Moreno-Betancur, M. and Carlin, J. B. (2018). Understanding Interventional Effects: A More Natural Approach to Mediation Analysis? *Epidemiology*, 29(5):614–617.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, Cambridge.

Neumark, D. and Rich, J. (2019). Do Field Experiments on Labor and Housing Markets Overstate Discrimination? A Re-examination of the Evidence. *ILR Review*, 72(1):223–252.

Nguyen, T. Q., Ogburn, E. L., Schmid, I., Sarker, E. B., Greifer, N., Koning, I. M., and Stuart, E. A. (2023). Causal Mediation Analysis: From Simple to More Robust Strategies for Estimation of Marginal Natural (In)direct Effects. *Statistics Surveys*, 17:1–41.

Nguyen, T. Q., Schmid, I., Ogburn, E. L., and Stuart, E. A. (2022). Clarifying Causal Mediation Analysis: Effect Identification via Three Assumptions and Five Potential Outcomes. *Journal of Causal Inference*, 10(1):246–279.

Omi, M. and Winant, H. (2014). *Racial Formation in the United States, 3rd Edition*. Routledge, New York, NY.

Pager, D. (2003). The Mark of a Criminal Record. *American Journal of Sociology*, 108(5):937–975.

Park, S. and Esterling, K. M. (2021). Sensitivity Analysis for Pretreatment Confounding With Multiple Mediators. *Journal of Educational and Behavioral Statistics*, 46(1):85–108.

Paul, K. I. and Moser, K. (2009). Unemployment Impairs Mental Health: Meta-Analyses. *Journal of Vocational Behavior*, 74(3):264–282.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669–688.

Pearl, J. (2001). Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420, San Francisco, CA. Morgan Kaufmann Publishers Inc.

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

Pearl, J. (2010). The Foundations of Causal Inference. *Sociological Methodology*, 40(1):75–149.

Portes, A. and Rumbaut, R. G. (2006). *Immigrant America: A Portrait*. University of California Press, Berekley, CA.

Qin, X. (2024). Sample Size and Power Calculations for Causal Mediation Analysis: A Tutorial and Shiny App. *Behavior Research Methods*, 56:1738–1769.

Qin, X., Deutsch, J., and Hong, G. (2021). Unpacking Complex Mediation Mechanisms and Their Heterogeneity between Sites in a Job Corps Evaluation. *Journal of Policy Analysis and Management*, 40(1):158–190.

Qin, X. and Hong, G. (2017). A Weighting Method for Assessing Between-Site Heterogeneity in Causal Mediation Mechanism. *Journal of Educational and Behavioral Statistics*, 42(3):308–340.

Qin, X., Hong, G., Deutsch, J., and Bein, E. (2019). Multisite Causal Mediation Analysis in the Presence of Complex Sample and Survey Designs and Non-Random Non-Response. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(4):1343–1370.

Quillian, L. and Lee, J. J. (2023). Trends in Racial and Ethnic Discrimination in Hiring in Six Western Countries. *Proceedings of the National Academy of Sciences*, 120(6):e2212875120.

Quillian, L. and Midtboen, A. H. (2021). Comparative Perspectives on Racial Discrimination in Hiring: The Rise of Field Experiments. *Annual Review of Sociology*, 47:391–415.

Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, 1(3):385–401.

Rao, J. N. and Wu, C. (1988). Resampling Inference With Complex Survey Data. *Journal of the American Statistical Association*, 83(401):231–241.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE Publications, Thousand Oaks, CA.

Raudenbush, S. W. and Schwartz, D. (2020). Randomized Experiments in Education, with Implications for Multilevel Causal Inference. *Annual Review of Statistics and Its Application*, 7(1):177–208.

Rendall, M. S., Weden, M. M., Favreault, M. M., and Waldron, H. (2011). The Protective Effect of Marriage for Survival: A Review and Update. *Demography*, 48(2):481–506.

Ridley, M., Rao, G., Schilbach, F., and Patel, V. (2020). Poverty, Depression, and Anxiety: Causal Evidence and Mechanisms. *Science*, 370(6522):eaay0214.

Robins, J. (1986). A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period-Application to Control of the Healthy Worker Survivor Effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Robins, J. M. and Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, 3(2):143–155.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5):550–560.

Roelfs, D. J., Shor, E., Davidson, K. W., and Schwartz, J. E. (2011). Losing Life and Livelihood: A Systematic Review and Meta-Analysis of Unemployment and All-Cause Mortality. *Social Science & Medicine*, 72(6):840–854.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: the Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. C. (2005). Autobiographical Memory Tasks in Cognitive Research. In Wenzel, A. and Rubin, D. C., editors, *Cognitive Methods and Their Application to Clinical Research*, pages 219–241. American Psychological Association.

Samoilenko, M. and Lefebvre, G. (2019). Risk Ratio Equations for Natural Direct and Indirect Effects in Causal Mediation Analysis of a Binary Mediator and a Binary Outcome. *American Journal of Epidemiology*, 188(7):1201–1203.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Scheve, K. F. and Slaughter, M. J. (2001). *Globalization and the Perceptions of American Workers.* Institute for International Economics, Washington, DC.

Schoemann, A. M., Boulton, A. J., and Short, S. D. (2017). Determining Power and Sample Size for Simple and Complex Mediation Models. *Social Psychological and Personality Science*, 8(4):379–386.

Shields-Zeeman, L. and Smit, F. (2022). The Impact of Income on Mental Health. *The Lancet Public Health*, 7(6):E486–E487.

Shultz, T. R. (1982). Causal Reasoning in the Social and Nonsocial Realms. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 14(4):307–322.

Silverstone, P. H. and Salsali, M. (2003). The Relationship between Low Self-Esteem and Psychiatric Diagnosis. *Annals of General Hospital Psychiatry*, 2(2):1–9.

Simon, R. J. and Lynch, J. P. (1999). A Comparative Assessment of Public Opinion toward Immigrants and Immigration Policies. *International Migration Review*, 33(2):455–467.

Sniderman, P. M., Peri, P., de Figueiredo, Rui J.P., J., and Piazza, T. P. (2000). *The Outsider: Prejudice and Politics in Italy.* Princeton University Press, Princeton, NJ.

Sobel, M. E. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, 13:290–312.

Sobel, M. E. (2008). Identification of Causal Parameters in Randomized Studies With Mediating Variables. *Journal of Educational and Behavioral Statistics*, 33(2):230–251.

Tai, A.-S., Lin, S.-H., Chu, Y.-C., Yu, T., Puhan, M. A., and VanderWeele, T. (2023). Causal Mediation Analysis with Multiple Time-varying Mediators. *Epidemiology*, 34(1):8–19.

Tchetgen Tchetgen, E. J. (2013). Inverse Odds Ratio-Weighted Estimation for Causal Mediation Analysis. *Statistics in Medicine*, 32(26):4567–4580.

Tchetgen Tchetgen, E. J. (2014). A Note on Formulae for Causal Mediation Analysis in an Odds Ratio Context. *Epidemiologic methods*, 2(1):21–31.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric Theory for Causal Mediation Analysis: Efficiency Bounds, Multiple Robustness, and Sensitivity Analysis. *Annals of Statistics*, 40(3):1816–1845.

Tein, J.-Y. and MacKinnon, D. P. (2003). Estimating Mediated Effects with Survival Data. In *New Developments in Psychometrics: Proceedings of the International Meeting of the Psychometric Society*, pages 405–412. Springer.

Thomson, R. M., Igelstrom, E., Purba, A. K., Shimonovich, M., Thomson, H., McCartney, G., Reeves, A., Leyland, A., Pearce, A., and Katikireddi, S. V. (2022). How Do Income Changes Impact on Mental Health and Wellbeing for Working-Age Adults? A Systematic Review and Meta-Analysis. *The Lancet Public Health*, 7(6):E515–E528.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Trinh, M., Yamamoto, T., and Zhou, X. (2021). paths: An Imputation Approach to Estimating Path-Specific Causal Effects.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* Springer Series in Statistics. Springer Publishing, New York, NY.

Umberson, D. and Karas Montez, J. (2010). Social Relationships and Health: A Flashpoint for Health Policy. *Journal of Health and Social Behavior*, 51:S54–S66.

Valeri, L., Lin, X., and VanderWeele, T. J. (2014). Mediation Analysis when a Continuous Mediator is Measured with Error and the Outcome Follows a Generalized Linear Model. *Statistics in Medicine*, 33(28):4875–4890.

Valeri, L. and VanderWeele, T. J. (2013). Mediation Analysis Allowing for Exposure-Mediator Interactions and Causal Interpretation: Theoretical Assumptions and Implementation with SAS and SPSS Macros. *Psychological Methods*, 18(2):137–150.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer Series in Statistics. Springer Publishing, New York, NY.

VanderWeele, T. and Vansteelandt, S. (2014). Mediation Analysis with Multiple Mediators. *Epidemiologic Methods*, 2(1):95–115.

VanderWeele, T. J. (2009a). Concerning the Consistency Assumption in Causal Inference. *Epidemiology*, 20(6):880–883.

VanderWeele, T. J. (2009b). Marginal Structural Models for the Estimation of Direct and Indirect Effects. *Epidemiology*, 20(1):18–26.

VanderWeele, T. J. (2010). Bias Formulas for Sensitivity Analysis for Direct and Indirect Effects. *Epidemiology*, 21(4):540–551.

VanderWeele, T. J. (2011a). Causal Mediation Analysis with Survival Data. *Epidemiology*, 22(4):582–585.

VanderWeele, T. J. (2011b). Controlled Direct and Mediated Effects: Definition, Identification and Bounds. *Scandinavian Journal of Statistics*, 38(3):551–563.

VanderWeele, T. J. (2013). A Three-way Decomposition of a Total Effect into Direct, Indirect, and Interactive Effects. *Epidemiology*, 24(2):224–232.

VanderWeele, T. J. (2014). A Unification of Mediation and Interaction: A Four-Way Decomposition. *Epidemiology*, 25(5):749–761.

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction.* Oxford University Press, New York, NY.

VanderWeele, T. J. (2020). Frontiers of Power Assessment in Mediation Analysis. *American Journal of Epidemiology*, 189(12):1568–1570.

VanderWeele, T. J. and Arah, O. A. (2011). Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders. *Epidemiology*, 22(1):42–52.

VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation Analysis with Time-Varying Mediators and Exposures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):917–938.

VanderWeele, T. J., Valeri, L., and Ogburn, E. L. (2012). The Role of Measurement Error and Misclassification in Mediation Analysis. *Epidemiology*, 23(4):561–564.

VanderWeele, T. J. and Vansteelandt, S. (2010). Odds Ratios for Mediation Analysis for a Dichotomous Outcome. *American Journal of Epidemiology*, 172(12):1339–1348.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect Decomposition in the Presence of an Exposure-induced Mediator-outcome Confounder. *Epidemiology*, 25(2):300–306.

Vansteelandt, S., Bekaert, M., and Lange, T. (2012). Imputation Strategies for the Estimation of Natural Direct and Indirect Effects. *Epidemiologic Methods*, 1(1):131–158.

Vansteelandt, S. and Daniel, R. M. (2017). Interventional Effects for Mediation Analysis with Multiple Mediators. *Epidemiology*, 28(2):258–265.

Vansteelandt, S. and VanderWeele, T. J. (2012). Natural Direct and Indirect Effects on the Exposed: Effect Decomposition under Weaker Assumptions. *Biometrics*, 68(4):1019–1027.

Vinokur, A. D., Price, R. H., and Schul, Y. (1995). Impact of the JOBS Intervention on Unemployed Workers Varying in Risk for Depression. *American Journal of Community Psychology*, 23(1):39–74.

Vinokur, A. D. and Schul, Y. (1997). Mastery and Inoculation Against Setbacks as Active Ingredients in the JOBS Intervention for the Unemployed. *Journal of Consulting and Clinical Psychology*, 65(5):867–877.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wager, S. and Walther, G. (2015). Adaptive Concentration of Regression Trees, with Application to Random Forests. *arXiv preprint arXiv:1503.06388*.

Wang, W. and Albert, J. M. (2012). Estimation of Mediation Effects for Zero-Inflated Regression Models. *Statistics in Medicine*, 31(26):3118–3132.

Warren, J. R. (2009). Socioeconomic Status and Health across the Life Course: A Test of the Social Causation and Health Selection Hypotheses. *Social Forces*, 87(4):2125–2153.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133.

Wehenkel, A. and Louppe, G. (2021). Graphical Normalizing Flows. In *International Conference on Artificial Intelligence and Statistics*, pages 37–45. PMLR.

Wing, C., Simon, K., and Bello-Gomez, R. A. (2018). Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. *Annual Review of Public Health*, 39:453–469.

Wodtke, G. T. (2013). Duration and Timing of Exposure to Neighborhood Poverty and the Risk of Adolescent Parenthood. *Demography*, 50:1765–1788.

Wodtke, G. T. (2020). Regression-based Adjustment for Time-varying Confounders. *Sociological Methods & Research*, 49(4):906–946.

Wodtke, G. T., Alaca, Z., and Zhou, X. (2020). Regression-With-Residuals Estimation of Marginal Effects: A Method of Adjusting for Treatment-Induced Confounders That may also be Effect Modifiers. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(1):311–332.

Wodtke, G. T. and Almirall, D. (2017). Estimating Moderated Causal Effects with Time-Varying Treatments and Time-Varying Moderators: Structural Nested Mean Models and Regression with Residuals. *Sociological Methodology*, 47(1):212–245.

Wodtke, G. T., Yildirim, U., Harding, D. J., and Elwert, F. (2023). Are Neighborhood Effects Explained by Differences in School Quality? *American Journal of Sociology*, 128(5):1472–1528.

Wodtke, G. T. and Zhou, X. (2020). Effect Decomposition in the Presence of Treatment-induced Confounding: A Regression-with-residuals Approach. *Epidemiology*, 31(3):369–375.

Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach.* Cengage, Boston, MA.

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research*, 20(7):557–585.

Wright, S. (1934). The Method of Path Coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215.

Xu, Y. (2017). Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis*, 25(1):57–76.

Yan, Y. and Williams, D. R. (1999). Socioeconomic Status and Mental Health. In Aneshensel, C. S. and Phelan, J., editors, *Handbook of the Sociology of Mental Health*, pages 151–166. Springer Publishing, New York, NY.

Zhang, Z. (2014). Monte Carlo Based Statistical Power Analysis for Mediation Models: Methods and Software. *Behavior Research Methods*, 46:1184–1198.

Zheng, W. and van der Laan, M. J. (2012). Targeted Maximum Likelihood Estimation of Natural Direct Effects. *The International Journal of Biostatistics*, 8(1).

Zheng, W. and van der Laan, M. J. (2017). Longitudinal Mediation Analysis with Time-varying Mediators and Exposures, with Application to Survival Outcomes. *Journal of Causal Inference*, 5(2):20160006.

Zhou, J. and Wodtke, G. T. (2024). Causal Mediation Analysis with Multiple Mediators: A Simulation Approach. *unpublished manuscript.*

Zhou, X. (2020). Some Doubly and Multiply Robust Estimators of Controlled Direct Effects. *arXiv preprint arXiv:2011.09569.*

Zhou, X. (2022). Semiparametric Estimation for Causal Mediation Analysis with Multiple Causally Ordered Mediators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):794–821.

Zhou, X. and Wodtke, G. T. (2019). A Regression-with-Residuals Method for Estimating Controlled Direct Effects. *Political Analysis*, 27(3):360–369.

Zhou, X. and Yamamoto, T. (2022). Tracing Causal Paths from Experimental and Observational Data. *The Journal of Politics*, 85(1):250–265.