UNIVERSITY OF CALIFORNIA

Los Angeles

Selection into the Sample and into Treatment: Tools for Internally Valid Causal Inference

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Adam Robert Rohde

2023

ABSTRACT OF THE DISSERTATION

Selection into the Sample and into Treatment:
Tools for Internally Valid Causal Inference

by

Adam Robert Rohde
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2023
Professor Chad J. Hazlett, Chair

Studies often seek to estimate the causal effect of a treatment on an outcome using a sample that has been drawn from a larger population. Such samples may not be randomly drawn and researcher may not observe all confounding variables that drive selection into treatment. Though researchers may be content to study causal effects averaged over only the sample in hand, selective sampling and unobserved confounding can still bias such effect estimates, threatening their "internal validity" (Campbell, 1957). Sample selection and unobserved confounding are related in how they threaten internal validity and can be examined in conjunction. We develop graphical tools to help evaluate threats from sample selection and unobserved confounding. These tools also allow us to determine when covariate adjustment can overcome these threats. It will not always be possible to solve these problems, however, and we thus generalize sensitivity analyses for unobserved confounding to also address sample selection. We then consider the use of instrumental variables, which in some cases can be biased by these concerns, but in other cases offer a solution to them. Finally, shifting emphasis to unobserved confounding, we discuss the use of placebo variables in partial identification.

The dissertation of Adam Robert Rohde is approved.

Thomas R. Belin

Jennie E. Brand

Onyebuchi A. Arah

Chad J. Hazlett, Committee Chair

University of California, Los Angeles

2023

To my wife . . .

who—among so many other things—

supported me unwaveringly throughout my time at UCLA.

I am forever grateful.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I acknowledge my advisor Chad Hazlett. His guidance, encouragement, and support have been more than I could have hoped for from an advisor. I am deeply appreciative of his willingness to take me on as his student and to create an environment in which I could pursue my interests while learning from his example. The field of causal inference is fortunate to have him. I would also like to acknowledge Onyebuchi Arah, Jennie Brand, and Thomas Belin. I am grateful for their advice, perspectives, and discussion. Thank you all for making my experience at UCLA enjoyable and productive. I would also like to acknowledge my classmates and friends from the statistics department and the causology reading group. Learning with you has made this experience a joy. Thank you also to Sandra Chan and Robert Murphy, who have been enduring mentors since well before I started at UCLA. Finally, my deepest gratitude to my wife, without whose support I could never have made it this far, and to my parents and brother who have always encouraged me.

CURRICULUM VITAE

2009–2013	B.A. in Mathematics and Economics, Boston College.
2012–2019	Consulting Associate, Charles River Associates.
2019–2022	C.Phil. in Statistics, University of California, Los Angeles.
2019–	Ph.D. Student in Statistics, University of California, Los Angeles.

PUBLICATIONS

Murphy, R. and Rohde, A. (2018). Rational bias in inflation expectations. *Eastern Economic Journal*, 44:153–171

CHAPTER 1

Introduction

Researchers often seek to estimate the causal effect of a treatment on an outcome using a sample that has been drawn from a larger population. Often such samples are not randomly drawn and researchers do not observe all confounding variables that drive selection into treatment. Though researchers may be content to study causal effects averaged over only the sample in hand, selective sampling can still bias estimates of even these causal effects, thus threatening the "internal validity" (Campbell, 1957) of effect estimates. These threats are closely related to, and can be examined in conjunction with, the problem of unobserved confounding in the selected sample. In this dissertation, we aim to help understand and overcome the threats that sample selection and unobserved confounding pose to internal validity. We start with a focus on sample selection.

That selective sampling can threaten even internal validity has long been known, and over the decades different research traditions have offered guidelines for assessing the threats to internal validity posed by sample selection. Further, it is not possible to know what the causal effect would be in any other population of eventual interest if we cannot first obtain an unbiased estimate in the observed sample—a result we formalize. In Chapter 2, we develop graphical tools to help clarify and evaluate threats from sample selection (together with possible additional sources of unobserved confounding). These tools also allow us to determine when covariate adjustment can overcome these threats. We employ formal graphical tools for causal reasoning to more fully and rigorously characterize the (i) the settings in which selective sampling does and does not bias the "internal effect estimate," and (ii) the conditions under which this bias can theoretically be corrected, and how to do so using covariate adjustment. These results are collectively conveyed through a graphical criterion that investigators can apply in their circumstances to examine the threats of bias and opportunities for correction given a graphical causal model. A number of common lessons emerge, including that many forms of sample selection, including selection processes influenced by the treatment or a mediator, are not always problematic. That said, the central lesson is that many complications may arise, requiring the researcher to use these tools to examine how selection processes may bias results and whether this can be corrected under specific causal structures the user cannot reject as implausible.

In many cases it will not be possible to confidently solve these problems by adjusting for observed covariates, and we thus introduce a sensitivity analysis framework that combines sample selection and (other) unobserved confounding threats. In Chapter 3, we discuss the omitted variable based sensitivity analyses of Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) and how these can be generalized to evaluate threats from sample selection and threats from unobserved confounding. Since sample selection as a threat to internal validity is typically the result of collider stratification, the parameters in naive applications of such sensitivity analyses can be difficult to interpret. We show how more interpretable expressions for the sensitivity parameters in these frameworks can be derived in some simple, parametric settings. Using these as a guide, we propose bounds on the parameters for the general, non-parametric settings by drawing on information theory.

Could additional information like an instrumental variable be used to overcome threats to internal validity from sample selection? How does sample selection alter the use and usability of instrumental variables? In Chapter 4, we discuss interesting cases in which sample selection presents opportunities to use instrumental variables and in which using instrumental variables can be used to overcome sample selection. However, we discuss how these opportunities may arise only in very specific settings, as is the case for credible instrumental variables in general. We also discuss the numerous threats that sample selection can pose to the credibility of instrumental variable approaches. To facilitate this discussion, we revise existing graphical criteria for instrumental variables to highlight the special role that sample selection plays in the instrumental variables setting. We do this by first introducing an extension to typical causal graphs that visualizes how sample selection alters the relationships between variables in the sample. We then provide rules (graphical criteria) that allow researchers to use these extended graphs to evaluate the key assumptions of instrumental variables in their own applications, while responsibly accounting for sample selection. In this way, we generalize recent discussions of sample selection and instrumental variables and connect these to existing graphical criteria.

Another way additional information can be used to estimate causal effects comes in the form of "placebo" or proxy variables. Shifting our focus from sample selection to unobserved confounding, in Chapter 5, we discuss how, in the quest to make defensible causal claims from observational data, it is sometimes possible to leverage information from "placebo treatments" and "placebo outcomes" (aka "negative control outcomes") unaffected by the treatment. Traditional approaches employing such information focus largely on point identification and assume (i) "perfect placebos" (placebo treatments have precisely zero effect on the outcome; the real treatment has precisely zero effect on a placebo outcome); and (ii) that the placebo treatment/outcome suffers the same amount of confounding as does the real treatment-outcome relationship on some scale ("equiconfounding"). We take a different approach, showing how the analysis of "omitted variable bias" in regression provides a flexible and powerful way to leverage information from placebo treatments and outcomes while violating these assumptions by postulated degrees. This allows investigators to examine results under (i) any chosen assumption about the relative strengths of confounding suffered by a placebo treatment/outcome compared to the true treatment-outcome relationship, and (ii) "imperfect placebos", i.e. placebo treatments with up to some postulated non-zero effect on the outcome, or placebo outcomes experiencing effects of treatment up to some postulated strength. These tools can be easily applied with any choice of placebo treatment or outcome. Pre-treatment outcomes can often be employed as placebo outcomes in this way. Relatedly, conventional difference-in-difference approaches, in both the repeated cross-section and panel settings, are a special case of this setting in which a strict equiconfounding assumption (regarding the pre-treatment outcome's relationship to the treatment group indicator) is claimed. We demonstrate multiple relaxations of this that are natural under our framework. In certain settings, these tools can also be applied to sample selection as a threat to internal validity.

Thus, in this dissertation, we explore the threats posed by sample selection and unobserved confounding to internal validity of causal effect estimates and discuss various approaches to overcome these threats. While our emphasis shifts, each tool we present can apply to both sample selection and unobserved confounding as threats to internal validity. I hope that the tools developed here aid researchers in recognizing, understanding, and taming threats to internal validity.

CHAPTER 2

Sample selection as a threat to internal validity

In applied quantitative research, we often estimate quantities of interest using samples of data drawn in non-random ways from a population of eventual scientific or policy interest. In other settings, we may have a sample that is representative of a sub-population of interest, whether by design or circumstance, but the sub-population may not be representative of a larger population from which the sample was drawn. Where we are interested in learning the causal effect of one thing (D) on another (Y), this can create two types of problems. One problem involves *external validity* (Campbell, 1957): how might the effect of interest look when averaged over some population of interest, as compared to its average in our sample? This is an important question which has achieved considerable recent attention (for a recent review and development, see Egami and Hartman (2022)). While investigators are accustomed to obviating this problem by restricting their inferences to the observed sample, the second and more pernicious problem is that selective sampling can threaten even *internal validity*, biasing our estimate of the treatment effect as defined as an average over the sample. This has long been known. For example, Berk (1983) states, that under selective sampling, "Both internal and external validity are implicated. There is no escape by limiting one's causal conclusions to the population from which the nonrandom sample was drawn (or even the sample itself)."

Notwithstanding increased recent attention to external validity through a rapidly advancing literature on "transportation" and "generalization" (e.g., Pearl and Bareinboim (2011); Pearl (2015b); Bareinboim and Pearl (2016); Correa et al. (2018, 2019); Egami and Hartman (2021)), internal validity remains a key concern for investigators for two reasons. First, internal validity may reasonably be of sufficient scientific interest in many settings. For example, we may study a causal mechanism that is thought to be approximately the same in most or all individuals or populations thereof. Or, we may design a study so that the sample in hand is already representative of a sufficiently interesting target group, e.g. those who would be eligible for a new policy or therapeutic intervention. Second, even where ultimate interest lies in a claim of generalization to some broader or different population, we formally show how internal validity is always required for (and easier to achieve than) external validity: identification of internally valid causal effects invokes a subset of the causal assumptions invoked for identification of causal effect estimates that generalize from the sample.[1]

What detailed knowledge must investigators have to avoid or remedy these biases? What types of selection processes are problematic and what types are not? Can the internal validity of randomized experiments be threatened by sample selection? Where there are threats, under what conditions can they be remedied and how? How can we deal with the problems of confounding and sample selection simultaneously? Methodologists in a variety of disciplines have long sought to address these perennial questions, providing careful examinations, guidelines, and methodological fixes since at least Campbell (1957), with well-known later contributions in Greenland (1977), Heckman (1979), Berk (1983), Hernán et al. (2004), and Elwert and Winship (2014), among others. However, these approaches do not amount to a comprehensive and rigorous treatment of the problem, nor have they resulted in an approach that a researcher can apply to reveal any problem or solution that exists for any causal structure they deem plausible in their setting. As noted by Berk (1983),

[1] While we offer a formalization of these claims, we do so to add rigor to statements that have long been made. For example, Campbell (1957) states that “Internal validity is the prior and indispensable consideration”, and Campbell and Stanley (1963) argue that “Internal validity is the basic minimum without which any experiment is uninterpretable...” Shadish et al. (2002) clarify that the primacy of internal validity is specific to “cause-probing research,” which is the context of our paper. Additionally, any claim of internal validity entails some conception of the population from which the sample was drawn. Without this, we cannot begin to analyze whether the sample selection mechanism threatens our ability to draw causal inferences for the selected sample, since the selection must be from some population.

"while considerable effort has been devoted to documenting sampling biases within traditional survey sample approaches...we are a very long way from a formal theory."

Fortunately, we are now in a position to answer Berk's 1983 call for such a formal theory, and indeed to provide a procedure for answering such questions as they apply to any causal structure. In recent years, researchers have benefited from the growth and formalization of methods that rigorously define causal quantities, characterize when they can or cannot be estimated from the data, and point to solutions for correcting biases, employing the devices of potential outcomes, structural causal models, and graphical causal models.[2] With these tools comes the possibility of more completely and rigorously posing and answering questions about sample selection and internal validity. However, existing graphical approaches neglect the role that sample selection plays for internal validity. Thus, we develop a graphical approach to understanding the threats such sample selection can pose for internally valid causal effect estimates. This involves first introducing "internal selection graphs", an extension of standard graphical approaches that visually shows the consequences of sample selection for the relationships between variables. Second, we provide rules for how to use these extended graphs to determine when causal quantities are identifiable under selective sampling. These tools aim to provide a wider audience with the ability to analyze how sample selection might threaten internal validity in their applications. Applied to a number of common causal structures, our tools support a few broad findings of note, including: (i) sample selection is not always problematic to internal validity (e.g., post-treatment selection, or confounders of selection and the outcome (Hernán, 2017) are not biasing on their own), (ii) some causal effects can still be identified when sample selection is based on a mediator,[3] (iii) sample selection can influence the identification of causal effects even when it is not a collider, and

[2] While we direct readers to texts such as Pearl (1988, 2009); Imbens and Rubin (2015); Hernán and Robins (2020) for a fuller review of these concepts, our goal is to explain our use of these tools as they arise so as to provide a mostly self-contained guide to users.

[3] Mediators are nodes in causal graphs the lie on at least one causal path from the treatment to the outcome. See below for a discussion of causal graphs and paths.

(iv) the threats from sample selection for internal validity are not the same as those for external validity, nor are the means of addressing those threats. While communicating those broad conclusions helps to signal the complexity and possibly non-intuitive nature of this problem, our key message echoing Berk, 1983 and Greenland, 2022, is that for any specific application, the details of the causal structure and sample selection mechanism determine whether sample selection threatens internal validity and what might be done about it. Our primary contribution is thus the graphical criterion we provide that enables investigators to reliably perform such diagnostic and prescriptive analyses in their setting.

2.1 Working example: Racial bias in policing

For concreteness, we employ a single working example throughout the paper.[4] Inspired by Knox et al. (2020), we look to data from police "stops" (an encounter in which a police officer stops and interacts with a civilian, on foot or in a vehicle). The question is then what can be learned from such data about how the *police-perceived* race of the civilian stopped alters the chances that police employ force in that encounter. The emphasis on *police-perceived* race is important for two reasons. First, it reminds us of the possible misperception and conceptual ambiguity regarding the police officer's belief about the civilian, as opposed to how the civilian would identify. Second, it reminds us that we are interested in the question of how the police officer's *belief* regarding the civilian's race might have influenced the outcome.[5]

The key challenge we consider here is that the data are limited to administrative records that are produced only when the police officer stops a civilian, thus making a report, citation, or arrest that appears in the data. Hence, such studies are restricted to a sample of civilian-

[4] Appendix A.8 offers additional illuminating—and perhaps entertaining—exercises which the reader can use to test and develop their understanding. These include understanding why taller NBA players are worse free-throw shooters, whether imagining applying eyeliner helps one lose weight, seeing if doing more can feel like less, and more.

[5] For additional discussion of perceived race and racism in causal studies see Grogger and Ridgeway (2006); Greiner and Rubin (2011); Robinson and Bailey (2019); Khazanchi et al. (2020); Lett et al. (2022).

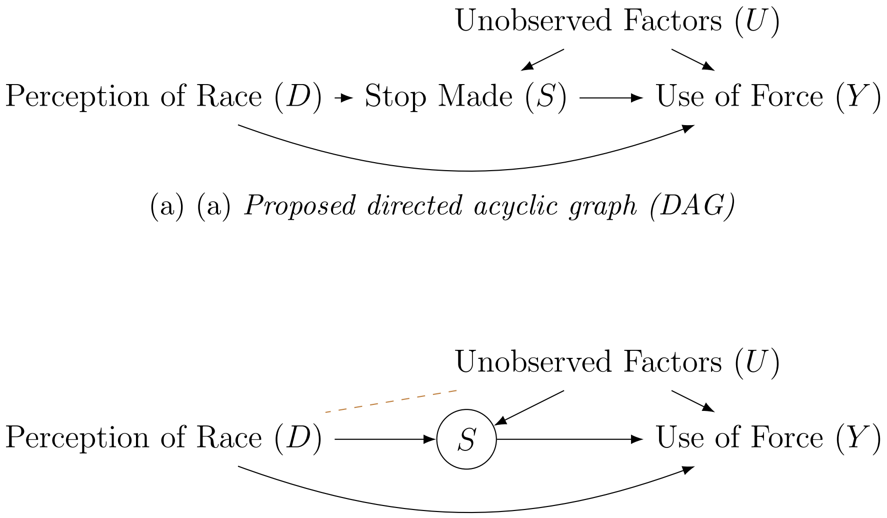
police encounters that has been selected in a non-random way, as the encounters in which the officers stop the civilian depends greatly on characteristics of the encounter (in particular on characteristics of the officer and the civilian).

To begin addressing this, we must first be willing to contemplate the causal structure of the system in question, meaning that we consider how each variable in this system *could* cause or be caused by other observed variables, or by a web of unobserved variables that influence more than one observable. Here we assume it is possible that police-perceived race influences the ways in which the officers interact with the civilian, both through whether or not the officers make a stop and whether or not the officers use force. Second, police stopping a civilian is a prerequisite to police use of force; if no stop is made, then officers cannot use force. Third, the police administrative records do not capture all of the factors that influence whether or not officers make a stop and/or use force; that is, there are unobserved common causes of making a stop and of using force. These possible relationships are represented in the usual causal graphical form (that is, as a directed acyclic graph (DAG)) in Figure 2.1(a), which mimics Knox et al. (2020) Figure 1.[6] Here D represents an indicator for police–civilian encounters involving a civilian that was perceived to be from a minority ethnic group, S represents an indicator for police–civilian encounters in which the police make a stop, Y is an indicator for police use of force, and U are the unobserved common causes of making a stop and use of force.

This causal structure immediately points to a number of familiar problems. We use Figure 2.1(b) to annotate the original DAG in ways that make these problems apparent. First, and central to this project, we are forced to "select" data for which a stop occurred

[6] For those unaccustomed to relying on DAGs, we note that it is important that the user include on such a DAG any arrow that could exist, meaning that to leave out an arrow requires a strong argument for why no such arrow exists. The requirement that such a causal structure be assumed at this level of detail may seem like a drawback. However, it asks little more than what is absolutely necessary to draw conclusions about how selection impacts the result and what can be done about it indeed depends on the causal structure at this level of detail. An unwillingness to transparently state what causal structure the researcher believes to be plausible would not free us from the consequences of that structure, only blind us to the possible problems and solutions associated with that structure.

Figure 2.1: Racial Discrimination in the Use of Force by Police



(a) (a) *Proposed directed acyclic graph (DAG)*

(b) (b) *Extended DAG revealing "bridge" formed by selection*

($S = 1$), and this stop is a mediator between perceived ethnicity (D) and the use of force (Y). This conditioning on S is represented by the circle around S in Figure 2.1(b), and it has two immediate consequences. First, because some of the effect of D flows through S to Y, conditioning on S in this way blocks some of the effect we wish to study, leaving only the part of the effect that flowed directly from D to Y as estimable. Second, by conditioning on $S = 1$, we are conditioning on a "collider" or a common consequence of D and U.[7] Conditioning on a collider (or a descendant[8] of a collider) *can* create purely statistical associations between the parents of the collider. We can represent this purely statistical association with a dashed undirected edge $D \cdots U$ on Figure 2.1 (b). This leads to additional problems. For example, by generating an association between D and U, the pathway $D \cdots U \rightarrow Y$ now generates an association between D and Y that is not due to the effect of D on Y. Thus, a comparison of the rates of use of force across values for (perceived) civilian race will be biased for the total

[7] A collider is a node in the graph into which two arrows point: $D \rightarrow S \leftarrow U$. See Pearl (2009) for an introduction to causal graphical models and colliders.

[8] We do not consider a node to be a descendant of itself. See Definition A.2.9 in Appendix A.

effect and direct effect *even for encounters in which a stop was made* (the selected sample). Knox et al. (2020) agree with these conclusions but use other tools to reach them.

We note that conditioning on a collider does not *always* create an association among its parents nodes, as there are a number of counterexamples. These counterexamples often impose strong or knife-edge assumptions and cannot easily be defended, but in other cases they are plausible. In Section A.6 of Appendix A, we discuss conditions under which marginally independent parents of a collider maintain their independence conditional on the collider for discrete variables and the implausible circumstances under which no association would be created between D and U conditional on $S = 1$ in the present example. In Section A.7 of Appendix A, we show that zero "interaction information" implies that marginally independent parents of a collider maintain their independence conditional on the collider. Because conditioning on a collider can and often will create an association among the parents, we proceed as though it does to maintain a more conservative analysis.

Finally, an important feature of these scenarios is to be clear and explicit about the causal estimand we are aiming to identify and estimate. In plain language, we are interested in "the effect of perceived race on use-of-force", specifically among (averaged over) those cases where a stop occurred. This description, however, is ambiguous regarding which counterfactuals we mean to compare. Specifically we could be referring to two different quantities:

- **The direct effect, among those who were stopped.** Once a stop is made, how does perceived race influence the risk that force is used? This considers an individual i for whom a stop was actually made, and compares whether force would have been used had they been perceived to be of one race ($D_i = d$) or another ($D_i = d'$).
- **The total effect, among those who were stopped** considers the "the effect of perceived race through triggering a stop that might not have otherwise occurred", as well as the direct effect above that applies "once a stop occurs". Consider an individual i with perceived race $D_i = d$, for whom a stop really did occur ($S_i = 1$). For this individual, had they been perceived to be of a different race ($D_i = d'$), a stop may or

may not have still occurred, and force may or may not have been used. We are interested in the outcome (use-of-force) for this individual, had we changed their perceived race, recognizing that doing so might also have changed whether a stop would have occurred. The average total effect among those who were stopped — or more generically the (average) "internal total effect" — is be composed of such counterfactual comparisons for all units i that were, in the real data, actually subject to a stop (without manipulation of their perceived race).[9]

2.2 Background

Notation and key quantities Let's introduce some notation to clarify the types of casual effects we mean when we say internally valid causal effects and causal quantities. A potential outcome, Y_{di}, is the value that the variable Y would have taken for unit i, if the variable D for unit i had been set, possibly counterfactually, to the value d. (Splawa-Neyman et al., 1990; Rubin, 1974, 1978, 1990) The unit-level causal effect of setting D to d relative to D to d' is $\tau_i = Y_{di} - Y_{d'i}$.

The fundamental problem of causal inference, however, is that we are never able to observe more than one of the potential outcomes for a given unit and so cannot calculate unit level

[9] Describing these quantities in terms of the "ideal" or "target" randomized trials can be a useful exercise, in part for the gap it may reveal between causal quantities and what can straightforwardly be learned by experimentation. For the "total effect," we must be able to randomly manipulate the perceived race of the civilian at the beginning of the encounter (before the officer decides whether to make a stop). Yet, to limit our inference to the individuals who would have been stopped, we must have a way of recording whether, when the officer perceived the race as they would have without intervention, they would have made a stop. Such an experiment would require a time-travel or memory-wiping mechanism, or the ability to monitor "two worlds" (one with the natural perceived race, one with the counterfactual). For the direct effect, the manipulation in question comes later: we wait and see if a stop is made, and only wipe the officer's memory and randomly intervene on their perception of race when they made a stop naturally. Alternatively we could intervene on perceived race prior to the encounter, determine when a stop would have occurred for unit i under their naturally perceived race (keeping only individuals i for whom that is the case), and then, when we set perceived race for each such individual i to its non-unobserved value, we must simultaneously force a stop to occur, even if it would not have otherwise. Such interventions are clearly infeasible, which gives some researchers pause, whereas others are satisfied to make such comparisons regardless, as they can be clearly defined by their counterfactuals and/or by the structural causal model and DAG they reference.

causal effects. (Rubin, 1978; Holland, 1986; Imbens and Rubin, 2015; Westreich et al., 2015) Despite this, these are often the building blocks of typical causal inferential targets. When readers see "internally valid causal effects," we suspect that most have in mind something like the sample average treatment effect (SATE), $\frac{1}{N}\sum_{i=1}^{N}\tau_i$, which is the simple average of the unit level effects across the units that are observed in the sample. This is a perfectly good target causal effect and the discussion that follows will apply to this. However, researchers might also be interested in the causal effect for the *sub-population for which the selected sample is representative.* We write this as $\mathbb{E}[\tau_i|S_i = 1]$ and call it the selected-population average treatment effect (SPATE). We will use a binary variable, S, to denote non-random sample selection. $S_i = 1$ can be interpreted as indicating that a unit is included in the observed study sample. It may also be thought of as indicating the subset of the population from which the observed study sample was randomly drawn.[10] Both the SATE and SPATE are different from the population average treatment effect (PATE), $\mathbb{E}[\tau_i]$.

$$
\begin{aligned}
\text{SATE} &= \frac{1}{N}\sum_{i=1}^{N}\tau_i \\
\text{SPATE} &= \mathbb{E}[\tau_i|S_i = 1] \\
\text{PATE} &= \mathbb{E}[\tau_i] = \mathbb{E}[\tau_i|S_i = 1]p(S_i = 1) + \mathbb{E}[\tau_i|S_i = 0]p(S_i = 0)
\end{aligned}
$$

In the case of a non-random sample from the population, we have that $\mathbb{E}[\text{SATE}|S_i = 1] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N}\tau_i|S_i = 1] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\tau_i|S_i = 1] = \mathbb{E}[\tau_i|S_i = 1] = \text{SPATE}$. As we discuss further below, $\text{SPATE} = \mathbb{E}[\tau_i|S_i = 1] = \mathbb{E}[\tau_i] = \text{PATE}$ when $Y_d \perp\!\!\!\perp S$, as in the case of a random

[10] In a setting like our running example, police-civilian encounters in the study sample are selected based on whether or not the police made a stop, which we represent with $S_i = 1$. The specific set of encounters that we observe and are able to analyze may not contain all encounters for which police made stops, however. Perhaps we only have a sample of such encounters - a random sample or one based on geography, time, or other constraints. If we are willing to assume that these other constraints are not material to the causal relationships between perceived race and use of force, then we might aim to estimate a causal effect for all police-civilian encounters in which a stop was made. The fact that we only have a sample of all such encounters does not pose a problem for causal identification of causal effects for the sub-population of encounters in which a stop was made. However, only having a sample that contains encounters in which a stop was made may threaten our ability to identify such a causal effect.

sample from the population. In the non-random sampling case, the SATE is unbiased for the SPATE, not the PATE, since the sample at hand can be thought of as a representative sample of the sub-population indicated by $S = 1$.

This brings us to a more formal definition of internal validity: An estimation strategy is said to be "internally valid" if it can unbiasedly or consistently estimate the SPATE (of other causal effects for the selected population). In what follows, we do not always differentiate between units eligible to be in the selected sample from those specifically in the sample in hand. Obtaining a valid estimate of a causal effect for the specific sample, we can then generalize this to the subpopulation. So going forward, we often refer to just the units in the sample at hand, even if our target is really the subpopulation. Other causal effects (i.e., not just means of τ_i) might be of interest as well, but in what follows, we will focus on distributional primitives that allow for the identification of any causal effect.

Identification under conditional ignorability. Any internally valid causal effect can be written using $p(Y_d|S = 1)$. For example, $\mathbb{E}[Y_d - Y_{d'}|S = 1] = \sum y \times p(Y_d = y|S = 1) - \sum y \times p(Y_{d'} = y|S = 1)$. Such quantities can be identified under sufficiently strong assumptions, including that covariates Z are observed such that in each stratum of Z in the selected population, treatment is ignorable, i.e. $Y_d \perp\!\!\!\perp D|Z, S = 1$.[11]

$$\begin{aligned}
p(Y_d|S = 1) &= \sum_z p(Y_d|Z = z, S = 1)p(Z = z|S = 1) && \text{by law of total probability} \\
&= \sum_z p(Y_d|D = d, Z = z, S = 1)p(Z = z|S = 1) && \text{by } Y_d \perp\!\!\!\perp D|Z, S = 1 \\
&= \sum_z p(Y|D = d, Z = z, S = 1)p(Z = z|S = 1) && \text{by consistency}
\end{aligned}$$

Structural causal models and DAGs. The key practical concern is: how can we know if D and Y_d are in fact independent conditional on Z in the selected sub-population? It is

[11] Here, we ignore problems of measurement bias, missingness, interference, and other wrinkles that are important in practice.

simple to write a conditional ignorability statement of this kind, and even to understand a hypothetical randomization scheme that would justify such a choice. With observed data, however, supporting or defending such an assumption turns on detailed claims about the underlying data generating process. To convincingly defend an assumption of conditional ignorability is to convincingly argue that the data were generated from one of a class of "structural causal models" that would produce such ignorability. Specifically, such a model must make claims as to how the treatment and outcome causally relate to each other and relevant observed and unobserved covariates. The key assumptions in these models often regard which variables *do not* influence other variables. Such causal relationships can often be non-parametrically encoded in a structural causal model (SCM) which can be represented graphically as a directed acyclic graph (and extensions thereof). See Pearl (2009) and Appendix A for details on SCMs and DAGs. See Figure 2.2 for example DAGs. DAGs allow us to visualize dependencies and independencies between variables. On a given DAG, a set of rules known as graphical criteria can then be used to determine if ignorability holds under the sample selection mechanism implied by that DAG. However, existing approaches that deal with sample selection focus entirely on generalization and external validity. In addition, existing approaches focused on internal validity do not address sample selection explicitly and can be violated by sample selection in settings where causal progress is possible. In the following section, we therefore develop a graphical approach that incorporates sample selection for internal validity.

2.3 Proposal

Our set of tools for analyzing how sample selection can threaten internal validity begins with a method for displaying the effects of sample selection graphically. Next, we present a formal graphical criterion for determining whether internally valid causal effects are identifiable. Our proposed approach is similar to those presented in Pearl (1995); Shpitser et al. (2010);

Daniel et al. (2012); Correa et al. (2018), but emphasizes the impact of sample selection for internal validity regardless of the role that sample selection takes in the causal graph. We conclude this section by drawing connections with the generalizability results in Correa et al. (2018) and with a discussion of sample selection based on mediators in more detail. It is important to emphasize that we should have some population in mind from which the sample was selected. This will allow us to attempt to non-parametrically model the sample selection process.[12] As discussed above, we will use a binary variable, S, to denote sample selection.

2.3.1 Showing selection: internal selection graphs

We will now detail our simple graphical approach to determining whether conditional ignorability of the form $Y_d \perp\!\!\!\perp D | Z, S = 1$ holds. The key is to graphically represent the ways in which sample selection alters the relationships in the selected sample. We do this by defining internal selection graphs, which visually extend traditional causal graphs to represent all the ways that sample selection can change relationships between variables.[13]

Definition 1 (Internal Selection Graph, G_S^+)**.** Let G be the DAG induced by a SCM.

1. Create G_S by adding an appropriately connected binary selection node, S.
2. Draw a circle around S to clearly indicate that we must limit our analysis to $S = 1$.
3. Add to G_S any node which is a parent of the treatment or a parent of a descendant of the treatment. (U_S, the background factors contributing to selection, can be excluded.)
4. Add a dashed undirected edge between all variables between which S is a collider or an ancestor of S is a collider. We will call these dashed, undirected edges *bridges*.

Call the resulting graph an *internal selection graph, G_S^+*.

[12] While having a population in mind is important, “There is also the problem of infinite regress. Even if one has a random sample from a defined population, that population is almost certainly a nonrandom subset from a more general population. ... In principle, therefore, there exists an almost infinite regress for any dataset in which at some point sample selection bias becomes a potential problem.” Berk (1983)

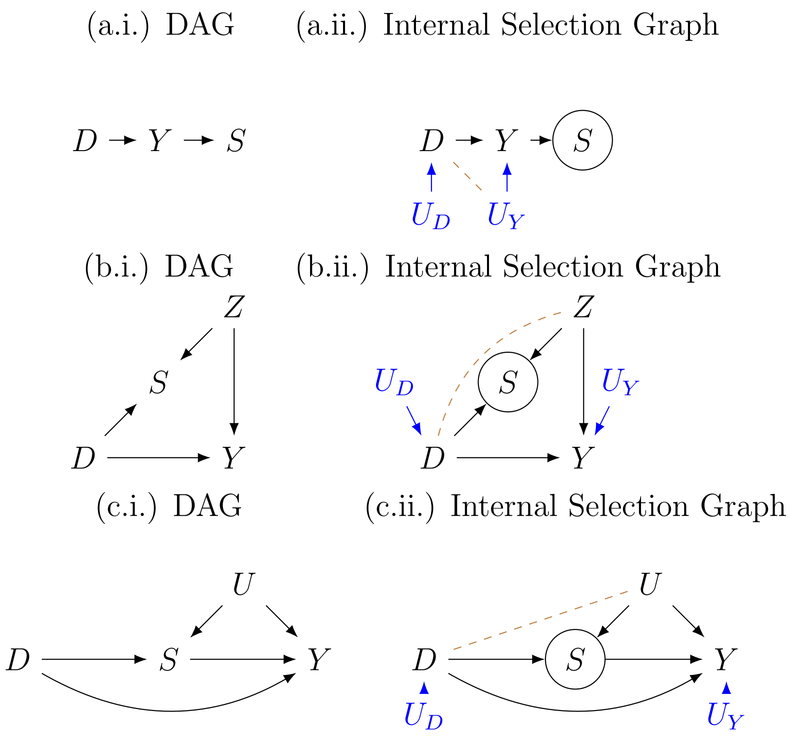
[13] See Appendix A for a brief discussion of why we do not use Single World Intervention Graphs. (Richardson and Robins, 2013a)

(This definition is similar to the "modified extended diagram" in Daniel et al. (2012).)

Let us dwell briefly on why we've chosen to denote sample selection with a separate binary selection node, rather than conditioning on some variable already in the causal model. Consider a simple two node DAG, $D \to Y$, in which both variables are continuous and there is sample selection on the outcome. Suppose this represents the effect of education (D) on income (Y) in a simplified setting in which we assume no common causes of education and income. (Elwert and Winship, 2014; Hausman and Wise, 1977) If we have a sample that is filtered to units with a particular *range* of (low) incomes, then sample selection, S, is not identical to income. There is still some variation in income, despite sample selection. As such, we should represent sample selection as a child of income: $Y \to S$. In other cases, like our running example related to the the effect of racial discrimination on police use of force, selection may be equivalent to a binary variable in the causal model (in our example sample selection is equivalent to police making a stop). In this case, we should represent the existing binary variable as the sample selection variable. There are also many settings in which sample selection is caused by more than one variable and the binary representation can simplify things while also showing how such variables can become related by selection. In cases when the binary selection variable is not equivalent to another variable in the causal graph, selection may not eliminate all variation in that other variable. This means that paths running through that variable may not be blocked by sample selection. Only paths that run directly through variables equivalent to selection will be blocked. We are not the first to use a binary selection variable in this way. See Bareinboim et al. (2014); Correa et al. (2018); Egami and Hartman (2021), among others. We echo Greenland (2022) in the sentiment that "realistic causal diagrams should always have a selection (sampling) indicator node S ... as a part of the data-generating process."[14]

[14] This sentiment is not new, though the graphical form may be. Berk (1983) states "When considering whether potential sample selection bias is likely to be realized, the initial step is to formulate a theoretical model of the selection process. One needs a theory of selection. Without a theory, it is difficult to draw even preliminary inferences about the nature of the problem and impossible to choose how best to implement

Figure 2.2: Examples of DAGs and Internal Selection Graphs


(a.i.) DAG (a.ii.) Internal Selection Graph

(b.i.) DAG (b.ii.) Internal Selection Graph

(c.i.) DAG (c.ii.) Internal Selection Graph

The key features of internal selection graphs[15] are the inclusion of an encircled sample selection node, specific background variables,[16] and bridges[17] that capture the statistical

sample selection corrections.”

[15] The value of this sort of graph for evaluating sample selection can be seen in Elwert and Winship (2014); Schneider (2020), papers that explore various types of selection bias in sociology and economic history. These papers, without stating a formal approach for doing so, add bridge-like undirected edges to the graphs they use to illustrate issues related to sample selection, but do not formally discuss how these non-causal edges can be incorporated into attempts to identify causal quantities, as we do in this paper. Greenland et al. (1999); Daniel et al. (2012) also discuss approaches in which undirected edges are added to the graph.

[16] Step three of the internal selection graph definition requires that we include additional nodes in our causal graph. This means including background noise nodes for the treatment and all descendants of the treatment. As a result, these nodes are entirely determined by their parents represented in the graph. Including U_S would lead to the direct parents of S to be associated with each other via bridges through U_S. But the direct parents of S will already be associated with each other and connected with a bridge due to conditioning on selection itself. The associations between U_S and any direct parents of S are otherwise immaterial to ignorability, making the inclusion of U_S unnecessary.

[17] Bridges are simply graphical representations of the purely statistical relationships that arise as a result of conditioning on a collider or the purely statistical alteration of relationships as a result of sample selection. Here, we are forced to filter to $S = 1$; so when S is a collider, we are conditioning on a collider.

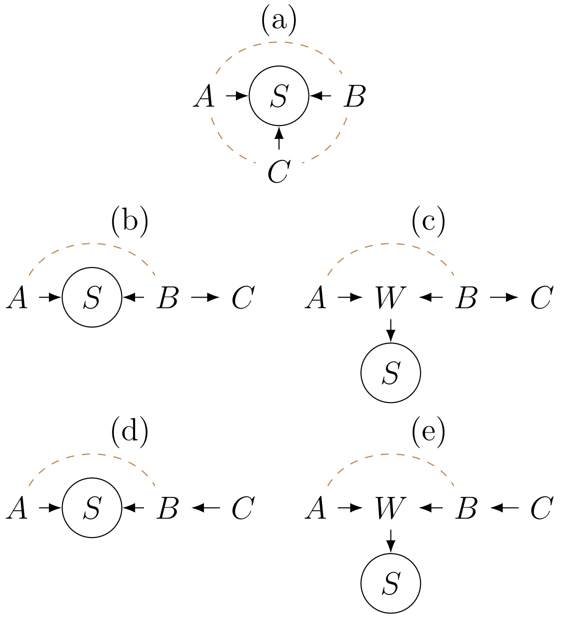
associations that result from sample selection. Bridges are created as a result of conditioning on a collider; this is also referred to as "collider stratification".[18] These additions ensure sample selection and the changes it requires for identification are visualized in the graph and can be analyzed easily. See Figure 2.2 for examples. Figure 2.2(c.ii.) contains the internal selection graph for our working example. At first glance, it appears that the types of sub-paths in internal selection graphs has expanded. We might wonder if the usual chains, forks, and colliders are joined by additional sub paths containing bridges. But the new additions are just built up from the old. The possible sub-paths incldue: chains ($A \to B \to C$), forks ($A \leftarrow B \to C$), colliders ($A \to B \leftarrow C$), bridge chains ($A \cdots B \to C$ or $A \cdots B \leftarrow C$) and double bridges ($A \cdots B \cdots C$). A double bridge might result from something like in Figure 2.3(a). A bridge chain might result from something like Figure 2.3(b,c,d,e). So colliders are still defined only with respect to directed edges. Bridges cannot create colliders, since they are really just graphical representations of purely statistical relationships created by sample selection.

Next, we will differentiate between a few types of paths. *Generalized paths* are any sequence of nodes and edges (directed edges and/or bridges) where each node appears only once (e.g., $D \cdots Z \to Y$, $D \to Y$, $D \to S \leftarrow Z$, $U_D \to D \to Y$). *Causal paths* are any generalized path where all edges between the nodes are directed and point in the same direction (e.g., $D \to Y$, $U_D \to D \to Y$). *Generalized non-causal paths* are any generalized path that isn't a causal path (e.g., $D \cdots Z \to Y$).[19] Any generalized non-causal path between D and Y can be considered a "confounding path" that provides an alternative explanation

[18] Conditioning on a collider allows information to flow between the direct parents of the collider. This is the opposite of what happens when we condition on other types of nodes, in which case conditioning "blocks" the flow of information along a path. We can state these relationships as follows. Two sets of nodes, D, Y, in a graph G are said to be *d-separated* by a third set, Z, if every path from any node $D_0 \in D$ to any node in $Y_0 \in Y$ is blocked. A path is blocked by Z if either [1] some W is a collider on the path between D, Y and $W \notin Z$ and the descendants of W are not in Z or [2] W is not a collider on the path but $W \in Z$ (see e.g. Pearl, 2009, Chapter 1 for details).

[19] Following the above discussion, d-separation is defined in the same way for these paths as for regular paths, since colliders are defined in the same way. See Appendix A, in particular, Corollary 5 and Definition A.2.10 for details.

Figure 2.3: New sub-paths are built from the old.



for why D and Y covary, other than due to their causal relationship. Considering causal paths (those that we would like to study) versus confounding paths (those we do not want to study) can be useful for building intuition for how graphical criteria work.

Consider again the two node example of education and income where the sample has been selected to include only low income individuals. Figure 2.2(a) captures this. We do not assume that education explains the entirety of the variation in income and add a separate selection node. We also add the U_Y background noise term. Since S is a descendant of Y and Y is a collider between D and U_Y, there is a purely statistical association created between D and U_Y. Therefore, there is a generalized non-causal path from $D \cdots U_Y \to Y$ that will confound estimates of the effect of education on income, despite the fact that we assumed there were no common causes between education and income in this simplified example. This is the simplest example of the need to include such background terms in the analysis of how sample selection alters relationships between variables in the sample. Let us also consider again our running example of racial discrimination's effect on police use of force for civilian-police encounters in which the police make a stop. This is described in Figure 2.2(c). We see that the internal selection graph is essentially identical to the graph in Figure 2.1(b) but with two additional background terms. In this example, the background terms do not play a role, but as we've just seen they can play an important role in many cases. They are included here since internal selection graphs are constructed so that all the ways that sample selection can alter relationships in the sample are captured for any causal graph and sample selection mechanism.

2.3.2 Internal Selection Adjustment Criterion

How can we use internal selection graphs to determine whether conditional ignorability holds more generally? Using the intuition of ruling out "alternative explanations" (other than the causal effect of interest) for why the treatment and outcome covary, we want to leave the causal paths between the treatment and the outcome that we hope to study untouched while removing any other systematic relationships between the treatment and outcome that could provide an alternate explanation. That is, first, we don't want to conditioning on variables that are on causal paths we are interested in, since this would block some of the effect we want to study. Second, we want to block any open generalized non-causal paths between the treatment and outcome. These paths are not those we want to study and confound the causal paths that we do want to study. Third, we don't want to open any previously closed generalized non-causal paths between treatment and outcome as a result of our covariate adjustment. (Pearl, 2009) Sample selection may block paths whether we would like them blocked or not and may open previously closed paths. We formalize these intuitions as a set of rules that fold in the effects of sample selection; we refer to them as the internal selection adjustment criterion.

Definition 2 (Internal Selection Adjustment Criterion (ISAC))**.** A set of nodes Z in G_S^+ satisfies the internal selection adjustment criterion relative to D (treatment) and Y (outcome) if

1. No element of Z lies on or is a descendant of a node that lies on a causal path originating from D and arriving at Y. Note that an element of Z could be a descendant of D itself, if it is not on a causal path from D to Y. Note also that elements of Z should not be on, or descendants of nodes on, causal paths even if S is also on the causal path.
2. Z blocks every generalized non-causal path between D and Y that does not pass through S. Note that generalized non-causal paths passing through M, when S is a descendant of mediator M, on which M is an ancestor of Y also do not need to be blocked, assuming the previous condition is not violated. Further, M should not be a member of Z from the previous condition.

An important component of this criterion is that we do not need to worry about blocking paths on which S appears. If S is a collider, we have already added bridges that circumvent S itself. If S is not a collider but is on a path, it blocks the path.[20] This also shows us that sample selection can alter relationships in the sample, even when the sample selection node is not a collider. As in the case when the sample selection node is a collider, we also need to take care to account for how it alters relationships when it is not a collider.

Our criterion builds on powerful, well-known, existing criteria like those presented in Pearl (1995); Shpitser et al. (2010); Daniel et al. (2012); Correa et al. (2018). Using these building blocks, ISAC focuses us on what is required to attain internal validity under non-random sample selection by relying on internal selection graphs and allowing us to determine whether

[20] When the sample selection node is on a generalized non-causal path but is not a collider, we do not need to consider blocking this path with some additional node Z. Sample selection will already block this path. Further, the associations created when sample selection is a collider are captured by the bridges that circumvent S. So we do not need to consider any path that passes through S. When selection is the descendant of a mediator M, we will be working with potential outcomes for which we intervene on M; see below. This means that paths running through M on which M is an ancestor of Y will be blocked and since we already condition on a descendant of M, parents of M will already be associated due to selection, when M is a collider.

causal effects are identifiable under non-random sample selection, regardless of the role that sample selection plays in the casual graph. See Section 2.4 for a more detailed discussion. The following results use our internal selection adjustment criterion to show how to identify internal causal quantities in the presence of sample selection and confounding, whether selection is post-treatment or not. Recall that mediators are nodes that lie on at least one causal path from the treatment to the outcome. Mediators are discussed further below.

Internal Validity Results. If a set of nodes Z in G_S^+ satisfies ISAC relative to D (treatment) and Y (outcome) and

- *S is not a mediator or descendant of a mediator between D and Y,* then
  - (ignorability) $Y_d \perp\!\!\!\perp D | Z, S = 1$.
  - (identification) We can identify $p(Y_d | S = 1) = \sum_z p(Y | D = d, Z = z, S = 1) p(Z = z | S = 1)$.
- *S is a mediator between D and Y (but S is not also a descendant of another mediator),* then
  - (ignorability) $Y_{d,S=1} \perp\!\!\!\perp D | Z, S = 1$.
  - (identification) We can identify $p(Y_{d,S=1} | S = 1) = \sum_z p(Y | D = d, Z = z, S = 1) p(Z = z | S = 1)$.
  - (non-ignorability) For any set of observables, W, $Y_d \not\perp\!\!\!\perp D | W, S = 1$ and $Y_{d,S_{d'}} \not\perp\!\!\!\perp D | W, S = 1$.
- *S is a descendant of an observed mediator, M, between D and Y (but S is not also a mediator itself),* then
  - (ignorability) $Y_{d,m} \perp\!\!\!\perp D | Z, S = 1$ and $Y_{d,m} \perp\!\!\!\perp M | D, Z, S = 1$, where $M = m$ is observed.
  - (identification) We can identify $p(Y_{d,m} | S = 1) = \sum_z p(Y | D = d, M = m, Z = z, S = 1) p(Z = z | S = 1)$.
  - (non-ignorability) For any set of observables, W, $Y_d \not\perp\!\!\!\perp D | W, S = 1$ and $Y_{d,M_{d'}} \not\perp\!\!\!\perp D | W, S = 1$.

These results are proved in Appendix A in Theorems 6, 7, and 8. Note that all of these identification results equate internal causal quantities with expressions that are estimable from the selected sample alone.[21] The first bullet can be used to identify "internal total effects"; the second and third bullets can be used to identify "internal direct effects." The above results show that sample selection can often be seen as an omitted variable problem. Heckman (1979) discusses "the bias that results from using nonrandomly selected samples to estimate behavioral relationships as an ordinary specification bias that arises because of a missing data problem." The idea is essentially that, in many cases, sample selection can be thought of as an omitted variable or misspecification problem. Heckman's discussion was in the context of a parametric framework and he proposed a correction procedure in this context. Our results show that something similar is true when we take the graphical, non-parametric view on sample selection as a threat to internal validity. When the threat that sample selection poses to the internal validity of causal effect estimates of D (treatment) on Y (outcome) can be overcome through adjustment on some covariates, Z, the problem of sample selection for internal validity can be viewed as an omitted variable problem, when Z is unobserved. Before turning to a discussion of internal direct effects and mediators, we first consider some connections to generalizability.

2.3.3 Connections to generalizability

As previously discussed, researchers are often concerned with how a causal effect might look averaged over some population of interest, as opposed to averaged over the sample at hand. When the sample is drawn from the target population in some (random or non-random) way, we may hope to use information on the causal effect available from the sample to

[21] Estimation strategies that would apply to a covariate adjustment or conditional ignorability identification strategy will also apply when we properly account for sample selection. In Section A.5 of Appendix A, we present a discussion of IPW estimation of $\mathbb{E}[Y_d|S=1]$, $\mathbb{E}[Y_{d,S=1}|S=1]$, and $\mathbb{E}[Y_{d,m}|S=1]$. This is meant to illustrate one example of how estimation might proceed for internal causal effects. We also provide this discussion to demonstrate that estimation of "internal controlled direct effects" is also straightforward.

generalize to a statement about the causal effect in the population.[22] Generalizability in this sense is a form of external validity. We will show that internal validity is more permissive than generalization, by which we mean the causal assumptions invoked to identify (and the observed data required to estimate) internally valid causal effects are a subset of those invoked for causal effect estimates that generalize from the sample. In doing so, we formalize what Campbell and Stanley (1963) first claimed with respect to experiments for approaches using covariate adjustment: "Internal validity is the basic minimum without which any experiment is uninterpretable..."[23][24] We illustrate this in more formal terms using graphical criteria; we limit our discussion to causal quantities containing potential outcomes of the form Y_d for simplicity, i.e. those relating to total effects.

Definition 3 (Generalization Criterion (GC))**.** A set of nodes Z in G_S^+ satisfies the generalization criterion relative to D (treatment) and Y (outcome) if

- Z satisfies ISAC relative to D and Y and
- $Z_{\text{Ext}} \subset Z$ blocks all causal and generalized non-causal paths between Y and S in G_S^+ other than those that end in a causal path from D to Y.

[22] In this section, we discussion connections between our results and recent results addressing generalization (Correa et al., 2018). This is a distinct but related to the problem of transportability. See Bareinboim and Pearl (2016) for a discussion of the differences between generalization and transportability. We are considering a single SCM. In many interesting observational settings, the same SCM might not hold across all settings of interest. For example, there may be a reason why, at the particular hospital we have data from, the SCM we are evaluating holds and further that we can sustain related ignorability statements or satisfy graphical criteria related to this SCM, but the SCM might be different at other hospitals. Often we have to be identification opportunists, looking for someplace that the treatment was assigned in some way that is conducive to identification, but those mechanisms (SCM) may not hold elsewhere. You might be able to extend to a larger population in which the same SCM holds, but not to a population in which the SCM does not hold. When considering external validity, you have to have a specific target population in mind; and here we're talking about external validity for the population in which the same SCM holds; i.e., generalization not transportation.

[23] Shadish et al. (2002) clarify that the special role of internal validity is specific to "cause-probing research", as we are discussing here, but not all forms of research.

[24] An alternative version of this idea can be stated as follows. If observed covariates and data limited to the selected sample ($S = 1$) are insufficient identify $p(Y_d|S = 1)$, then they will also be insufficient to identify $p(Y_d) = p(Y_d|S = 1)p(S = 1) + p(Y_d|S = 0)p(S = 0)$, since $p(Y_d)$ extends beyond the selected sample which is all we have information about. We show this within the context of graphical criteria in this section.

This definition is a translation of Definition 8 from Correa et al. (2018).

Generalization Results. If a set of nodes Z in G_S^+ satisfies GC relative to D (treatment) and Y (outcome) and S *is not a mediator or descendant of a mediator between* D *and* Y, then

- (ignorability) $Y_d \perp\!\!\!\perp D|Z, S = 1$ and $Y_d \perp\!\!\!\perp S|Z_{\text{Ext}}$.
- (identification) We can identify $p(Y_d) = \sum_z p(Y|D = d, Z = z, S = 1)p(Z_{\text{Int}} = z_{\text{Int}}|Z_{\text{Ext}} = z_{\text{Ext}}, S = 1)p(Z_{\text{Ext}} = z_{\text{Ext}})$, where $Z_{\text{Int}} = Z - Z_{\text{Ext}}$.

This translates Definition 7 and Theorem 1 in Correa et al. (2018) to use potential outcomes.

The causal assumptions invoked by ISAC to identify $p(Y_d|S = 1)$ are a subset of those invoked by GC to identify $p(Y_d)$ using covariate adjustment. Not only do we need $Y_d \perp\!\!\!\perp D|Z, S = 1$, we also need $Y_d \perp\!\!\!\perp S|Z_{\text{Ext}}$. Further, we can estimate $p(Y_d|S = 1)$ with observed data from the selected sample for Y, D, Z alone. On the other hand, $p(Y_d)$ can be estimated with observed data from the full population on Z_{Ext} in addition to data from the selected sample for Y, D, Z. In this way, we have formalized that generalization "costs more" in terms of both causal assumptions and observed data. These generalization results are proved in Appendix A; see Theorem 9. Additional connections are also shown in Lemma 24.[25]

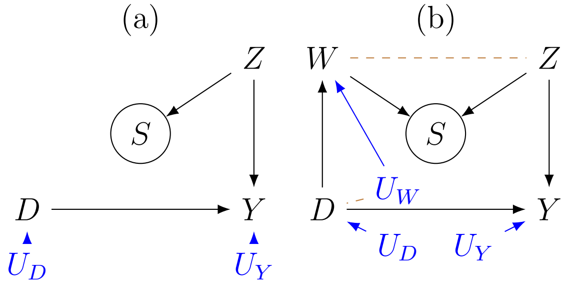
Let's explore some examples and discuss how identifiability can differ for internal and external validity. First, let's consider our running example of racial discrimination in policing in which we are interested in the effect of perceived civilian race on police use of force in which we have data only on civilian-police encounters in which the police made a stop. The internal selection graph for this can be found in Figure 2.2(c.ii.). We immediately see that there is no hope for generalizing: since S is a direct cause of Y there is no hope of satisfying GC, regardless of the choice of Z_{Ext}. Further, we only have access to data for police-civilian encounters in which the police made a stop, so we do not have any data for the full population. But do we satisfy the internal selection adjustment criterion? As we saw before, no we don't.

[25] See Correa et al. (2018) for an IPW estimator for generalization.

This is because of the generalized non-causal path from $D \cdots U \rightarrow Y$, which violates ISAC. So no total or direct effect is identifiable.

Next let's consider Figure 2.4(a), which we will just consider in the abstract. Here there is a common cause of sample selection and the outcome, but the two are not directly associated and sample selection is not a collider. Let's consider internal validity first. Suppose we choose not to condition on any covariates. We easily see that we satisfy ISAC. This means that we are indeed able to identify internal causal quantities without any covariate adjustment. We also only need data from the selected sample to estimate effects for the selected sample. What about identifying external causal quantities? We see that letting $Z = Z_{\text{Ext}}$ means that we satisfy GC. So we can identify external causal quantities. However, since we need to adjust for Z, we could only estimate generalized effects so long as we have data for Z for the full population. Finally, Figure 2.4(b) is similar to Figure 2.4(a), but it turns out that we can identify internal causal quantities adjusting for W or Z or both. Estimation would only require data on one of these for the selected sample. However, to identify external causal quantities we must adjust for Z (or Z and W). Estimation would require data on Z for the full population, data for the selected sample alone is insufficient, as is data only on W, even if it is for the entire population.

Figure 2.4: Examples for comparison of ISAC and GC.



2.3.4 Sample Selection based on a mediator

In this section, we return to potential outcomes of the the form $Y_{d,S=1}$ and $Y_{d,m}$, which allow us to consider particular types of direct effects called "controlled direct effects." Above we saw that, when sample selection plays the role of a mediator or the descendant of a mediator and we satisfy ISAC, we can identify $p(Y_{d,S=1}|S = 1)$ or $p(Y_{d,m}|S = 1)$, respectively. Identifying these internal causal quantities allows us to identify what we call "internal controlled direct effects," which are defined below. But what are these strange looking effects?

Definition 4 (Internal Controlled Direct Effects (ICDEs))**.** Define $\mathbb{E}[Y_{D=d,S=1} - Y_{D=d',S=1}|S = 1]$ and $\mathbb{E}[Y_{D=d,M=m} - Y_{D=d',M=m}|S = 1]$ to be the internal controlled direct effect when selection is a mediator between D and Y and when selection is a descendant of a mediator, M, between D and Y, respectively.

In mediation analysis, we have a treatment D, a mediator M, and an outcome Y, in addition to other relevant covariates. There are two possible paths along which the treatment might effect the outcome. First is the familiar direct path: $D \rightarrow Y$. Second is the indirect path: $D \rightarrow M \rightarrow Y$. This set up follows the mediation discussion from Baron and Kenny (1986). For our purposes, we consider the settings in which the sample selection node is itself a mediator between D and Y[26] or is a descendant of a mediator between D and Y. There are a variety of causal effects to consider when considering mediation, including total effects, controlled direct effects, natural direct effects, and natural indirect effects (Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2011; VanderWeele and Vansteelandt, 2009; Richiardi et al., 2013). In both of our settings, a causal path between D and Y is blocked or partially blocked as a result of sample selection. This means that total effects, like SPATE $= \mathbb{E}[Y_d - Y_{d'}|S = 1]$, which capture all causal paths along which the treatment D can effect the outcome Y, are not identifiable. The indirect effect of D on Y that runs along the

[26] Selection can be a mediator only when it is equivalent to a substantive binary variable in the causal graph. Our working example is an example of this.

path on which S lies or is a descendant will also not be identifiable. Only the causal paths $D \to Y$ that do not relate to S remain unaltered. Are we able to identify the direct effects that run along these unaltered paths?

In our present discussion, we are limited to the study sample. So the type of causal effects that we may be able to identify when sample selection is a mediator or descendant of a mediator will still be for the selected sample alone. We are therefore interested in direct effects for the selected sample. ICDEs are just this sort of effect: direct effects averaged over the units in the selected sample. ICDEs are distinct from CDEs for the entire population and are also distinct from total effects for the selected sample. Instead, ICDEs compare setting $D = d$ with setting $D = d'$, while also setting M to m (or S to 1) for both versions of treatment interventions, *for units in the selected study sample.* To understand ICDEs better, let's return to our working example. First, we note that total effects cannot be identified since sample selection (police making a stop) is a mediator blocking one of the causal paths between the treatment (police perception of race) and outcome (police use of force). But we might consider an ICDE as a possible effect of interest. The ICDE is the difference in police use of force between

- a setting in which we intervene to force police to perceive each civilian as being from the *minority* racial group and intervene to force police to stop each civilian
- a setting in which we intervene to force police to perceive each civilian as being from the *majority* racial group and intervene to force police to stop each civilian

where we average this difference in use of force only over the police-civilian encounters in which police actually did make a stop. Therefore, the ICDE in this example captures the effect of perceptions of rave on use of force, once a stop occurs, averaged over the ecnounters in which a stop was made in reality. A key to understanding this effect is to consider that, for some encounters in which a stop was made in reality, under a different perceived civilian race the officer may not have made a stop. ICDEs evaluate what would have happened if we intervened in these cases to ensure that the officer still made a stop. In this example, the path $D \cdots U \rightarrow Y$ violates ISAC and confounds even the ICDE. That is, sample selection creates a purely statistical association between perceptions of race and use of force that we cannot disentangle from the direct effect, even for police-civilian encounters we observe.

2.4 Discussion

2.4.1 Connections to existing work

Numerous literatures relate closely to the problems of sample selection and internal validity, yet we argue that the tools proposed here fill a gap in the toolkit available to researchers to fully and reliably examine the potential threats to internal validity due to sample selection and to illuminate possible solutions. Sample selection can arise at various points in the a study: during study entry (e.g., from non-participation or participation that is not representative of the population) or the data gathering process (e.g., only gathering data on some segment of the population), between study entry and analysis (e.g., loss to follow-up), or even during analysis as a result of conditioning or subsetting. Montgomery et al. (2018) illustrate how sample selection can threaten not only observational studies but also experiments. Sample selection and the associated bias goes by many different names in various fields: sample truncation bias, non-response bias, attrition bias, ascertainment bias, Heckman selection bias (Heckman, 1979), selection on the treatment, selection on the outcome, Berkson's bias, homophily bias, survival bias, m-bias, differential loss to follow up, volunteer bias, self-selection bias, healthy worker bias, and others. See Hernán et al. (2004); Elwert and Winship (2014); Schneider (2020) for overviews of the various forms that sample selection can take. Different research traditions have proposed informal guidelines for determining when sample selection threatens internal validity. Most notably, Campbell, Stanley, and their co-authors (Campbell, 1957; Campbell and Stanley, 1963; Cook and Campbell, 1979; Shadish et al., 2002) introduce the language of validity[27] to discuss conceptually how bias can arise in the design and implementation of studies and how these challenges can be overcome. Pearl and others (Pearl, 1995, 2009, 2014; Shpitser et al., 2010; Spirtes et al., 2000) advocate causal graphs and structural causal models that can be seen as subsuming the potential outcomes approach (Rubin, 1974, 1978, 1990) to causal inquiry. Matthay and Glymour (2020) take the very useful step of explicitly connecting the graphical approach to the Campbell tradition.

There are graphical approaches focused on identification of causal effects (Pearl, 1995; Shpitser et al., 2010) that do not directly discuss sample selection. There are many recently-developed approaches focused on generalizability and transportability (Bareinboim and Pearl, 2012, 2016; Correa and Bareinboim, 2017; Correa et al., 2018, 2019; Bareinboim et al., 2014; Lesko et al., 2017; Pearl, 2015b; Pearl and Bareinboim, 2011, 2014, 2019; Hartman et al., 2015; Egami and Hartman, 2021). There are graphical approaches focused on generalizing conditional causal effects from complete cases when there is missing data (Daniel et al., 2012), generalizing in the face of missing data and sample selection (Saadati and Tian, 2019), and generalizing from missing data alone (Mohan and Pearl, 2014, 2021).[28] Didelez et al. (2010) focus on outcome dependent sampling, on the causal odds ratio, and mostly on generalization; they also present some results on testing for the presence of conditional causal effects.

Our criterion builds on these powerful, existing graphical criteria, but focuses on and facilitates easy analysis of internal validity under non-random sample selection by visualizing the effects of sample selection and illustrating whether identification of internally valid causal effects is possible regardless of the role that the sample selection node plays in the causal graph. For instance, ISAC is similar to the back-door criterion of Pearl (1995). However, understanding that sample selection is a form of conditioning that could be contemplated on a DAG, simply including S in the adjustment set of the back-door criterion would violate the

[27] Shadish et al. (2002) describe "internal validity" as concerning whether the covariation between the treatment and outcome results from a causal relationship for the study sample.

[28] Missing data can be seen as a generalization of sample selection. (Saadati and Tian, 2019; Westreich, 2012; Howe et al., 2015)

back-door criterion when S is post-treatment, since the back-door criterion does not allow post-treatment conditioning. Thus, we would violate the back-door criterion in cases where we can make causal progress. Further, Pearl (2009) Section 11.3 discusses how background noise nodes that are often left off of DAGs can become important under certain types of conditioning. We ensure that, when relevant, such variables are included on internal selection graphs so that all effects of sample selection are represented graphically. Second, ISAC is also similar to the adjustment criterion of Shpitser et al. (2010). However, simply including S in the adjustment set of the adjustment criterion would violate the adjustment criterion when S is a mediator or a descendant of a mediator. This is because the adjustment criterion disallows conditioning on nodes that appear on causal paths or that are descendants of nodes on causal paths. Thus, we would violate the adjustment criterion in cases where we can make causal progress. Third, ISAC is similar to the generalized back-door criterion of Daniel et al. (2012), which is focused on generalization from complete cases when there is missing data but not internal validity. However, the generalized back-door criterion does not allow for any post-treatment adjustment and, in Daniel et al. (2012), is shown to only to identify Z-conditional causal quantities. Finally, ISAC is similar to the generalized adjustment criterion of Correa et al. (2018), but ISAC is less restrictive since it focuses on internal rather than external validity. Specifically, ISAC requires a subset of the causal assumptions required by the generalized adjustment criterion.

Despite existing approaches not focusing on sample selection and internal validity, sample selection is recognized as a threat to internal validity that is fundamentally different from common cause confounding (Hernán et al., 2004; Hernán and Robins, 2020; Infante-Rivard and Cusson, 2018; Matthay and Glymour, 2020; Smith, 2020; Elwert and Winship, 2014; Schneider, 2020). Other authors discuss sample selection in the context of external validity (Arah, 2019; Flanders and Ye, 2019; Thompson and Arah, 2014) and mention the threat to internal validity posed by sample selection (Berk, 1983; Tripepi et al., 2010; Cuddeback et al., 2004; Hernán and Robins, 2006; Larzelere et al., 2004; Smith and VanderWeele, 2019;

Westreich et al., 2018; Mathur, 2022). The celebrated approach presented in Heckman (1979) makes parametric assumptions and requires data for the entire population to estimate a model for the probability of selection. Hernán (2017) shows how sample selection can present different problems for generalization and external validity than it does for internal validity; we provide clarity on this in the general. Recent papers have also parsed specific types of sample selection using graphical tools. (Lu et al., 2022; Sjölander, 2023) Finally, recent literature on competing events (Stensrud et al., 2021, 2022) appeal to the notion of separable effects in examples like our running policing example.

2.4.2 Lessons

We now draw several important lessons from our discussion. First, researchers need to be careful about the details of the causal graph that they are studying. Including a sample selection node in every causal model and careful consideration of the sample selection mechanisms is required to determine the threat that sample selection poses to internal validity and what, if anything, might be done about it. Second, there is potential for users to intentionally or unintentionally favor one graph over another very similar graph in order to show that some causal quantity is identifiable. These are difficult but inherent problems in causal study and good-faith efforts to do credible causal inference should spend ample time defending the specific causal model being analyzed. Let's consider what other lessons we can take away from what we've seen.

- Informal applications of other graphical criteria to internal validity and sample selection can be misleading and should be avoided.
- Sample selection can influence identification, even when it is not a collider.
- When sample selection manifests as attrition or some other post-treatment type of selection, randomization of treatment assignment does not automatically ameliorate sample selection problems even for internal validity. Therefore, the discussion here is not limited to observational studies.

- Sample selection does not always threaten internal validity or external validity.
- Selection on the outcome is usually problematic.[29] However, association of the outcome and selection does not automatically present a problem. When a third variable causes both and the two are only indirectly related, there may be no problem.
- Post-treatment selection is typically not a problem on its own.
- Indirect association between selection and the outcome or selection and the treatment are typically not problems on their own but can be when they appear together.
- When selection is a mediator, causal quantities can be identified.

We again emphasize that these types of lessons do not apply universally. The specific causal quantities that may be identified, and whether identification of them is possible, depends on the causal model and which variables are observed. Whether selection is a collider, confounder, mediator, or indirectly related to variables of interest, ISAC provides clear guidance on identification. Formal tools like this should be the default mode for determining when and how internally valid causal effects can be identified. The target of causal studies is often a causal effect averaged across the study sample alone. The ability to estimate such an effect without bias is called internal validity. We've presented a simple graphical framework for dealing with sample selection that allows us to reliably attain internal validity for causal effects for the selected study sample. This framework allows sample selection to play any role in the original causal graph and provides clear guidance on which causal quantities are identifiable and what is required for identification through covariate adjustment. We've also seen examples and lessons that, we hope are prove useful for researchers as they attempt to obtain internally valid estimates of causal effects for their study sample.

[29] Though it can be allowed in certain circumstances, like case control studies in which the causal odds ratio is the target. See Daniel et al. (2012) section 4.5 "Missingness driven only by outcome." This and examples like (1.h.) highlight that rules of thumb like "selection on the outcome is biasing" might not correctly characterize all scenarios. Hence, while such general lessons can be useful guidance, it is important to consider the specifics of each application and use formal tools and systematic analysis.

CHAPTER 3

Sensitivity analysis

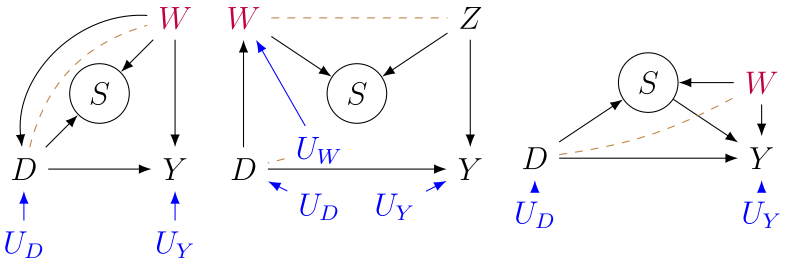
Both sample selection and unobserved confounding can threaten the internal validity of causal effect estimates by creating spurious associations (i.e., confounding paths) between the treatment and outcome variables in a dataset. In Chapter 2, we provide a framework for evaluating these threats and whether covariate adjustment can be used to eliminate them. However, there are many settings in which covariate adjustment on observed variables alone is insufficient to overcome these threats to internal validity. In such settings, sensitivity analysis based on the omitted variables paradigm can be productive. Existing sensitivity analysis frameworks (Cinelli and Hazlett, 2020; Chernozhukov et al., 2022) based on omitted variables can be generalized to address the threats to internal validity from sample selection, in addition to unobserved confounding. Some sensitivity frameworks have already been adapted to the sample selection and internal validity setting. However, when applied to sample selection and internal validity, the quantities that researchers are required to reason about (i.e. the sensitivity parameters) can be "not as intuitive to specify." (Smith and VanderWeele, 2019) Our main contribution is to develop an approach that formally bounds such unintuitive parameters in terms of quantities that are intuitive, thus allowing existing omitted variable based sensitivity analyses to be used for sample selection and unobserved confounding.

Suppose an investigator is interested in estimating the effect of a treatment, D, on an outcome, Y, for the selected sample alone or for the subpopulation for which the selected sample is a representative sample (denoted with $S = 1$).[1] Suppose further that the investigator

[1] See Chapter 2 for further discussion of internal validity, causal effects for the selected sample, and causal

knows or is willing to assume that $Y_d \not\perp\!\!\!\perp D|X, S = 1$ but $Y_d \perp\!\!\!\perp D|W, X, S = 1$ based on the tools discussed in Chapter 2. $Y_d \perp\!\!\!\perp D|W, X, S = 1$ is a conditional ignorability statement that can be used to identify internally valid casual effects. Y_{di} is a potential outcome; that is, value the variable Y would have taken for unit i, if the variable D for unit i had been set, possibly counterfactually, to the value d. Suppose that X contains observed covariates. If W is not observed, we can consider it as an omitted variable and use an omitted variable based sensitivity analysis to understand the threats to internal validity posed by sample selection. In the present setting, W may be a variable that blocks confounding paths or spurious associations between the treatment and outcome created by sample selection.[2] Figure 3.1 presents simple examples. The graphs in Figure 3.1 are internal selection graphs, which explicitly show how sample selection alters the relationships between variables in the selected sample. See Chapter 2 for how such graphs can be constructed.

Figure 3.1: Internal selection graph examples



In this Chapter we discuss the omitted variable based sensitivity analyses of Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) and how these can be leveraged to evaluate the threats that sample selection poses for internal validity. We discuss how the sensitivity parameters in these frameworks, in the sample selection setting, can be difficult to interpret. We show how alternative expressions for these sensitivity parameters can be

effects for the subpopulation for which the selected sample is representative.

[2] W could also be a simple common cause confounder of the Y, D relationship. This would put you in the settings already discussed in Cinelli and Hazlett (2020) and Chernozhukov et al. (2022).

derived in terms of more easily interpreted quantities in simple parametric settings. Using these parametric settings as inspiration, we then propose bounds on the difficult to interpret sensitivity parameters for general, non-parametric settings again in terms of more easily interpreted quantities.

3.1 Omitted variable based sensitivity analyses

We review three settings in which an investigator may consider the threats to internal validity from sample selection as a omitted variable bias problem. Each of these settings allow us to use expressions for or bounds on such bias as the basis for a sensitivity analysis. These settings are those considered in Cinelli and Hazlett (2020) and Chernozhukov et al. (2022).

Linear Model We may be interested in estimating a linear regression model, using the selected sample, like in Equation 3.1a. Since we know that $Y_d \not\perp\!\!\!\perp D|X, S = 1$ and $Y_d \perp\!\!\!\perp D|W, X, S = 1$, $\beta_{Y\sim D|X,S=1}$ contains some bias relative to what we would estimate if we were to include W in the regression, as in Equation 3.1b. We will refer to $\beta_{Y\sim D|X,S=1}$ as θ_s and $\beta_{Y\sim D|W,X,S=1}$ as θ_l. These are the parameters of interest for the "short" and "long" regressions, respectively.

$$\left[Y = \beta_{Y\sim D|X,S=1}D + X\beta_{Y\sim X|D,S=1} + \epsilon_s\right]|S = 1 \tag{3.1a}$$

$$\left[Y = \beta_{Y\sim D|W,X,S=1}D + X\beta_{Y\sim X|D,W,S=1} + \beta_{Y\sim W|D,X,S=1}W + \epsilon_l\right]|S = 1 \tag{3.1b}$$

Partially Linear Model Alternatively, we may be interested in estimating a partially linear model, as in Equation 3.2b. As in the fully linear case, we are only able to estimate Equation 3.2a, which omits W from our estimation.[3]

[3] In both the linear and partially linear cases, we assume that the user is considering how a linear or partially linear model (and the inclusion of a covariate in these models) differs from a fully non-parametric

$$[Y = \theta_s D + f_s(X) + \epsilon_s] \,|S = 1 \tag{3.2a}$$

$$[Y = \theta_l D + f_l(X, W) + \epsilon_l] \,|S = 1 \tag{3.2b}$$

Non-parametric Model Finally, we may be interested in estimating a linear functional of the conditional expectation function of the outcome in a non-parametric setting, like Equation 3.3b. Here we assume a binary treatment D. $Y_d = f_Y(d, X, W, U_Y)$ is the equation for Y in the structural causal model[4] under intervention to set $D = d$. Again, the investigator is only able to estimate Equation 3.3a, where $f^*_Y(D, X) \triangleq \mathbb{E}[Y|D, X] = \mathbb{E}[f_Y(D, X, W)|D, X]$.

$$\theta_s = \mathbb{E}[f^*_Y(1, X) - f^*_Y(0, X)|S = 1] \tag{3.3a}$$

$$\theta_l = \mathbb{E}[Y_1 - Y_0|S = 1] = \mathbb{E}[f_Y(1, X, W) - f_Y(0, X, W)|S = 1] \tag{3.3b}$$

In all three settings, there will be some bias resulting from not adjusting for W in our estimation. The bias is $\theta_s - \theta_l$. The bias may arise from sample selection or common cause confounding or both. We can use existing omitted variable bias frameworks to obtain expressions for the bias. Using these we can conduct sensitivity analysis to see how our estimate would change if we had included W in our estimation.

3.1.1 Expressions for omitted variable bias

For each of the settings above, Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) provide expressions for or bounds on the omitted variable bias that can be expressed in terms of simple sensitivity parameters that capture relationships between variables *in the sample.*

setting, before considering that they want to know how inclusion of W in the model changes θ_s.

[4] See Chapter 2 for more discussion of structural causal models under sample selection.

Linear Model Following Cinelli and Hazlett (2020), we can show that omitted variable bias for linear regression conditional on $S = 1$ can be expressed as in Equation 3.4. See Appendix B Section B.1 for the full derivation.

$$|\hat{\beta}_{Y\sim D|X,S=1} - \hat{\beta}_{Y\sim D|X,W,S=1}| = \frac{\text{SD}(Y^{\perp D,X}|S=1)}{\text{SD}(D^{\perp X}|S=1)} \sqrt{\frac{R^2_{Y\sim W|D,X,S=1} R^2_{W\sim D|X,S=1}}{1 - R^2_{W\sim D|X,S=1}}} \tag{3.4}$$

$\frac{\text{SD}(Y^{\perp D,X}|S=1)}{\text{SD}(D^{\perp X}|S=1)}$ is estimable from observed data. $R^2_{Y\sim W|D,X,S=1}$ is a partial R^2 that equals the fraction of residual variation in Y explained by W after partialling out both D and X, in the selected sample. $R^2_{W\sim D|X,S=1}$ is a partial R^2 that equals the fraction of the residual variation in D explained by W after partialling out X, in the selected sample. See Cinelli and Hazlett (2020) for further discussion of how to interpret partial R^2s.

Partially Linear Model Chernozhukov et al. (2022) show that omitted variable bias in the partially linear setting can be bounded by an expression in terms of $\eta^2_{Y\sim W|D,X,S=1}$, $\eta^2_{D\sim W|X,S=1}$, and terms that are estimable from the data. $\eta^2_{Y\sim W|D,X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ are Pearson's correlation ratios (or non-parametric R^2s).[5] $\eta^2_{Y\sim W|D,X,S=1}$ is the proportion of residual variation in Y explained by W, in the selected sample. $\eta^2_{D\sim W|X,S=1}$ is the proportion of residual variation in D explained by W, in the selected sample. See Chernozhukov et al. (2022) for the specific bound and further discussion of how to interpret partial η^2s.

Non-parametric Model Chernozhukov et al. (2022) also show that omitted variable bias in the non-parametric setting can be bounded by an expression in terms of $\eta^2_{Y\sim W|D,X,S=1}$ and a second term that, in the case of targeting $\theta_l = \mathbb{E}[Y_1 - Y_0|S = 1]$ with a binary treatment D, is the "average gain in the conditional precision with which we predict D by using W in

[5] $\eta^2_{D\sim W|X,S=1} = \frac{\text{Var}(\mathbb{E}[D|W,X,S=1]|S=1)-\text{Var}(\mathbb{E}[D|X,S=1]|S=1)}{\text{Var}(D|S=1)-\text{Var}(\mathbb{E}[D|X,S=1]|S=1)} = \frac{\eta^2_{D\sim WX|S=1}-\eta^2_{D\sim X|S=1}}{1-\eta^2_{D\sim X|S=1}}$. $\eta^2_{Y\sim W|D,X,S=1}$ can be similarly interpreted.

addition to X," which is somewhat similar to $\eta^2_{D\sim W|X,S=1}$. See Chernozhukov et al. (2022) for the specific bound and details.

The sensitivity parameters in the above settings are either partial R^2s or η^2s, measures of dependence with which most researchers will have some familiarity.[6] However, our ability to productively conduct sensitivity analysis hinges on our ability to *interpret* the sensitivity parameters on which the sensitivity analysis relies. Researchers intending to conduct a sensitivity analysis using these approaches must be able to build an understanding of the relationships captured by these sensitivity parameters based on external knowledge, first principles, existing literature, intuition, and subject matter expertise. Such understanding reflects causal relations between the variables.[7] As we discuss next, obtaining such knowledge is complicated by sample selection.

3.1.2 Difficulty interpreting sensitivity parameters in selected samples

Let us allow for either of the sensitivity parameters $R^2_{Y\sim W|D,X,S=1}$ or $R^2_{W\sim D|X,S=1}$ (or $\eta^2_{Y\sim W|D,X,S=1}$ or $\eta^2_{D\sim W|X,S=1}$) to contain a spurious relationship that results from stratifying to $S = 1$ where S, sample selection, is a collider,[8] as opposed to or in addition to causal relationships that operate in the full population. Sensitivity analysis is productive when researchers are able to leverage external knowledge about the relationships captured by the

[6] While these may be familiar measures of dependence, they do not have all properties of dependence measures that are desirable. For example, an R^2 of zero can exist for dependent variables and an R^2 reflects only the linear relationship between variables. Their familiarity and simplicity may compensate for some of their drawbacks, however. See Rényi (1959) for a discussion of these measures of dependence and the properties that make good measures of dependence. Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) provide thorough discussions of sensitivity analysis using these bias expressions and reasoning about these types of sensitivity parameters, in addition to tools and examples for conducting such analysis. We refer the reader to those discussions for important guidance on these measures and sensitivity analysis in general.

[7] It is important to recognize that the R^2s or η^2s in the expressions for or bounds on omitted variable bias from Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) are statistical measures of dependence between the variables. These could measure direct causal relationships between variables or more complex chains of causal relationships.

[8] Colliders are nodes in causal graphs into which two edges point. For example, $A \rightarrow B \leftarrow C$.

sensitivity parameters to inform the range of plausible values that the sensitivity parameters may take. This can then be used to determine how θ_s may change if W were included in the estimation using bias expressions or bounds. However, such external knowledge will be difficult to obtain when a sensitivity parameter contains a purely statistical (i.e., non-causal) relationship created by conditioning on a collider due to sample selection.[9] This is because the association captured by such a sensitivity parameter does not result from structural relationships in which one variable causes another. Instead, the association results from, or is changed by, the often counterintuitive phenomenon of conditioning on a common effect (a collider). Such associations do not exist in the population from which the sample was selected, do not reflect mechanisms that appear in nature, and will, therefore, be difficult to understand from first principles, previous studies (unless those studies suffered from similar sample selection), intuition, or subject matter expertise concerning mechanisms. See the worked example below for an example. In what follows, we consider how we might deal with this problem by appealing to relationships between the variables in the full population, as opposed to the selected sample.[10] Relationships in the population should reflect causal mechanisms and should, therefore, be easier to reason about. In our discussion, we focus on $R^2_{W\sim D|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$. Similar discussion could apply to $R^2_{Y\sim W|D,X,S=1}$ or $\eta^2_{Y\sim W|D,X,S=1}$.[11] We start by building a sense for how these sensitivity parameters can be expressed in terms of structural relationships between the variables in the full population in simple parametric settings.

[9] By "purely statistical relationship," we mean one that arises due to conditioning, as opposed to a relationship that exists causally in the population from which the sample was selected.

[10] If one or both of the sensitivity parameters does not contain a relationship altered by collider stratification, then the parameters for the selected sample may be the same as those for the population. An exception is when sample selection blocks a causal path that operates in the population but not in the selected sample. See Chapter 2 for more discussion.

[11] We do not fully address the non-parametric case, since not all of the sensitivity parameters can be expressed as R^2s or η^2s, but if the sample selection collider alters the association between Y and W, then our discussion will still apply.

Binary random variables We consider the case where W and D are binary. We assume that data are generated according to a simple collider graph: $D \rightarrow S \leftarrow W$. Assume $X = \{\emptyset\}$. In Appendix B Section B.2, we show that $R^2_{D\sim W|S=1}$ can be written in terms of six probabilities as shown in Equation 3.5.[12]

$$R^2_{D\sim W|S=1} = [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]^2 \times \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{\begin{pmatrix} [P_{S=1|11}P_{D=1} + P_{S=1|10}P_{D=0}][P_{S=1|01}P_{D=1} + P_{S=1|00}P_{D=0}]\times \\ [P_{S=1|11}P_{W=1} + P_{S=1|01}P_{W=0}][P_{S=1|10}P_{W=1} + P_{S=1|00}P_{W=0}] \end{pmatrix}} \tag{3.5}$$

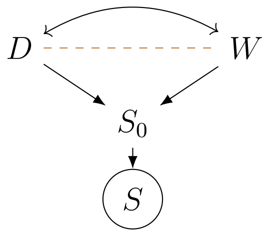
The relationship between W and D in the selected sample ($S = 1$) can be expressed in terms of the relationships between S, W and D, in the full population, in addition to $P(D = 1)$ and $P(W = 1)$. These quantities capture structural (i.e., causal) relationships between the variables. Thus, $R^2_{D\sim W|X,S=1}$ can be understood by considering structural relationships in the full population that comprise the sample selection mechanism.

Truncated multivariate normal random variables Next consider the case where W, D, and S have the causal structure shown in Figure 3.2, W, D, and S_0 have a multivariate normal joint distribution, and $S = \mathbf{1}[S_0 \geq \mathrm{C}]$ for some $\mathrm{C} \in \mathbb{R}$. Again $X = \{\emptyset\}$. S_0 can be thought of as a latent variable that captures how W and D relate to S. The bidirected edge captures that W and D could have some relationship in the population that is altered as a result of sample selection. Within the stratum $S = 1$, we have a truncated multivariate normal joint distribution. In Appendix B Section B.3, we show that $R^2_{W\sim D|S=1}$ can be expressed as in Equation 3.6.[13]

[12] $P_{W=w} = P(W = w)$, $P_{D=d} = P(D = d)$, $P_{S=1} = P(S = 1)$, $P_{S=1|wd} = P(S = 1|W = w, D = d)$, $P_{W=0} = 1 - P_{W=1}$ and $P_{D=0} = 1 - P_{D=1}$.

[13] In Equation 3.6, if $\rho_{W\sim D} = 0$, then $R^2_{D\sim W|S=1} = \frac{R^2_{S\sim D}R^2_{SW}\theta^2}{\sqrt{1-R^2_{S\sim D}\theta}\sqrt{1-R^2_{S\sim W}\theta}}$.

Figure 3.2: Internal selection graph for truncated multivariate normal example



$$R^2_{D\sim W|S=1} = \left(\frac{\rho_{D\sim W} - \rho_{S\sim D}\rho_{S\sim W}\theta}{\sqrt{1-\rho^2_{S\sim D}\theta}\sqrt{1-\rho^2_{S\sim W}\theta}} \right)^2, \tag{3.6}$$

where θ can be written as a function of $P(S=1)$ or C

The relationship between W and D in the selected (truncated) sample can be expressed in terms of the relationships between S, W and S, D as well as between W, D, in the full population. Additionally, we need $P(S=1)$, the probability of selection. Again, these quantities capture structural (i.e., causal) relationships that govern the sample selection mechanism and the relationship between W, D in the population.

Partial correlation and "constant selection effects" It is important to note that truncation on S or stratification to $S=1$ (i.e., sample selection) is not the same as linearly conditioning on S. Conditioning on S (linearly) would give Equation 3.7, based on the partial correlation formula.[14] This does not equal $R^2_{W\sim D|S=1}$ in general. Equations 3.6 and 3.7 are remarkably similar, however, with their only differences being the need to account for where truncation happens (i.e., the probability of selection). Equation 3.7 holds for linear conditioning on S, without any restrictions on the distribution or relationships between W, D, and S. Equation 3.6 only holds for truncated normals.

[14] In Equation 3.7, if $\rho_{W\sim D} = 0$, then $R^2_{D\sim W|S} = \frac{R^2_{S\sim D}R^2_{S\sim W}}{\sqrt{1-R^2_{S\sim D}}\sqrt{1-R^2_{S\sim W}}}$. Without restrictions on the distribution or relationships between W, D, and S, recall that $\rho_{WD} = 0$ does not mean that W and D are marginally independent.

$$R^2_{D\sim W|S} = \left(\frac{\rho_{D\sim W} - \rho_{S\sim D}\rho_{S\sim W}}{\sqrt{1-\rho^2_{S\sim D}}\sqrt{1-\rho^2_{S\sim W}}} \right)^2 \tag{3.7}$$

We might wonder under what circumstances we would be able to use Equation 3.7 to inform discussion of $R^2_{W\sim D|S=1}$. We explore this in Appendix B Section B.4. We could assume "constant selection effects" (akin to constant treatment effects) between the $S = 1$ and $S = 0$ strata. Such an approach would also require assumptions about the strata specific variances for W and D. While this could be used as a first pass analysis, the assumptions are typically unrealistic and using this approach could underestimate $R^2_{W\sim D|S=1}$.[15]

In the next section, we propose bounds on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ (as well as $R^2_{W\sim D|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$) for the non-parametric case in which we make no restrictions on the distribution or functional relationships between W, D, and S. In spirit, these bounds are similar to Equations 3.5 and 3.6 in that they ask us to reason about the relationships between W, D, and S (i.e., the sample selection mechanism) in the population, as well as the probability of selection. That we need to reason about the structural, causal relationships that define the sample selection mechanism in the population makes intuitive sense. This is the mechanism that leads to the spurious relationship in the selected sample in the first place. Researchers should be able to appeal to a combination of first principles, previous studies and existing literature, intuition, and subject matter expertise to understand the range of plausible strengths of these relationships.

3.2 Proposal

Since one of the sensitivity parameters in the omitted variable bias expression might contain some spurious association created by sample selection (and is therefore difficult to interpret),

[15] In Appendix B Section B.5, we discuss a simple bound on $R^2_{W\sim D|S=1}$ that relies on the partial correlation formula. However, this bound is typically uninformative (i.e., not less than 1).

we aim to bound this sensitivity parameter with structural relationships from the full population, about which investigators should be able to reason more easily. We will bound $R^2_{D\sim W|S=1}$, $\eta^2_{D\sim W|S=1}$, $R^2_{D\sim W|X,S=1}$, and $\eta^2_{D\sim W|X,S=1}$ by appealing to mutual information.

What is mutual information? Before we try to work with mutual information, what is it? Mutual information is a measure of how similar the joint distribution of two random variables, say A and B, is to the product of their marginal distributions. Therefore, it is a measurement of the total dependence between A and B, whether this dependence is linear or non-linear. It makes no assumptions about the distribution of A and B or the form their dependence takes. As we will discuss further below, it turns out that mutual information has a number of useful properties for measuring dependence that R^2s and η^2s do not. Mutual information between A and B, $\mathrm{MI}(A;B)$, can be thought of as the information obtained (or reduction in uncertainty) about variable A that results from learning the value of variable B. (Smith, 2015) Mutual information is defined in the following ways, where D_{KL} is Kullback–Leibler divergence and H is entropy.

$$\begin{aligned}\mathrm{MI}(A;B) &= D_{\mathrm{KL}}\left(P_{(A,B)} \| P_A \otimes P_B\right) \\ &= \sum_a \sum_b P_{(A,B)}(a,b) \log\left(\frac{P_{(A,B)}(a,b)}{P_A(a)P_B(b)}\right) = H(A) + H(B) - H(A,B)\end{aligned} \tag{3.8}$$

There are also useful notions of conditional mutual information and joint mutual information and entropy. See Ihara (1993); MacKay (2003); Cover and Thomas (2006) for details.[16] Mutual information measures the amount of Shannon information revealed about A as a result of knowing B. The Shannon information (or surprisal) of an event is defined as $I_A(a) = \log(1/P_A(a))$. Events that occur with certainty are perfectly unsurprising and hence

[16] We've shown the definition of mutual information for discrete random variables; analogous definitions exist for arbitrary random variables.

have no information. As the probability of an event decreases, the surprise that the event occurred increases, and so does the information content. The entropy of a random variable is the average information of the outcomes of the variable, $H(A) = \sum_a P_A(a) \log(1/P_A(a))$, and can be thought of as the uncertainty in the variable's outcomes. (MacKay, 2003) While mutual information can be an improvement as a measure of dependence over R^2 or η^2, in practice, interpreting mutual information can be difficult. Therefore, we appeal to a normalized version that has nice properties discussed below. Estimating mutual information and its normalized variants can also difficult in practice; however, we will be using it for sensitivity parameters that researchers reason about rather than estimate.

Mutual information for Gaussians In order to connect $R^2_{D\sim W|S=1}$, $\eta^2_{D\sim W|S=1}$, $R^2_{D\sim W|X,S=1}$, and $\eta^2_{D\sim W|X,S=1}$ with mutual information, we draw inspiration from the relationship between R^2 and mutual information for Gaussian random variables. For random variables, W and D, with a bivariate Gaussian joint distribution, there is an exact relationship between R^2 and mutual information. (Ihara, 1993; Cover and Thomas, 2006)

$$\mathrm{MI}(W; D) = -\frac{1}{2}\log(1 - R^2_{D\sim W}) \iff R^2_{D\sim W} = 1 - \exp(-2 \times \mathrm{MI}(W; D)) \tag{3.9}$$

While this relationship do not hold for arbitrary random variables, many authors have considered this type of transformation of mutual information as a way to obtain a useful mutual information based measure of dependence.[17] Lu (2011) presents such a measure of dependence that is defined for arbitrary variables and that has many nice properties we discuss below. We employ a slight variation on Lu (2011)'s L-measure, to create a useful normalized version of mutual information for our purposes. The L-measure takes the form $L(\mathrm{MI}) = 1 - \exp(-2 \times \mathrm{IF} \times \mathrm{MI})$, where IF is an "inflation factor" that ensures that the

[17] See Linfoot (1957); Kent (1983); Joe (1989); Kojadinovic (2005); Lu (2011); Speed (2011); Kinney and Atwal (2014); Asoodeh et al. (2015); Smith (2015); Shevlyakov and Vasilevskiy (2017); Laarne et al. (2021), among others.

L-measure takes appropriate values for arbitrary variables, not just continuous variables. See Appendix B Section B.6 for details.

Bounds This normalization of mutual information and Theorem 1 (below) allow us to build interpretable bounds on $R^2_{D\sim W|S=1}$, $\eta^2_{D\sim W|S=1}$, $R^2_{D\sim W|X,S=1}$, and $\eta^2_{D\sim W|X,S=1}$ without any assumptions on the distrubutions or functional relationships between the variables. Theorem 1 can be applied to the case for which S is a collider between D and W (e.g., $D \rightarrow S \leftarrow W$), demonstrating how conditioning on a collider alters the relationship between the parents of the collider. We leverage our normalized version of mutual information, which we call normalized scaled mutual information (NSMI), to give interpretable bounds on $R^2_{D\sim W|S=1}$, $\eta^2_{D\sim W|S=1}$, $R^2_{D\sim W|X,S=1}$, and $\eta^2_{D\sim W|X,S=1}$ that rely on Theorem 1. These bounds can be found in Theorems 2 and 3. These results are proved and NSMI is defined in detail in Appendix B Section B.6. While we use these results in the context of conditioning on a collider, S, they hold for stratification to $S = 1$ in general.

Theorem 1. *For random variables D, W, S, conditioning on S alters the relationship between D and W according to the expression $MI(D; W|S) = MI(D; W) + MI(S; [D, W]) - MI(S; D) - MI(S; W)$. Therefore, the change in dependence due to conditioning on S can be characterized using mutual information according to $MI(D; W|S) - MI(D; W) = MI(S; [D, W]) - MI(S; D) - MI(S; W)$. The dependence is not changed when $MI(S; [D, W]) = MI(S; D) + MI(S; W)$. When S is binary, it is also possible to write $MI(D; W|S) = p(S = 1)MI(D; W|S = 1) + p(S = 0)MI(D; W|S = 0)$, meaning that $MI(D; W|S = 1) \leq \frac{MI(D;W|S)}{p(S=1)} = \frac{MI(D;W)+MI(S;[D,W])-MI(D;S)-MI(W;S)}{p(S=1)}$.*

Theorem 2. *For random variables D, W, and S, where S is binary, the $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ resulting after stratification to $S = 1$ can be bounded in the following ways:*

1. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(S;[D,W])]}{[1-NSMI(S;D)][1-NSMI(S;W)]}\right)^{\frac{1}{p(S=1)}}$
2. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(D;S|W)]}{[1-NSMI(S;D)]}\right)^{\frac{1}{p(S=1)}}$
3. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(W;S|D)]}{[1-NSMI(S;W)]}\right)^{\frac{1}{p(S=1)}}$
4. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - ([1 - NSMI(D;W)][1 - NSMI(D;S|W)])^{\frac{1}{p(S=1)}}$
5. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - ([1 - NSMI(D;W)][1 - NSMI(W;S|D)])^{\frac{1}{p(S=1)}}$
6. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - ([1 - NSMI(D;W)][1 - NSMI(S;[D,W])])^{\frac{1}{p(S=1)}}$

Theorem 3. *For random variables D, W, S, X, where S is binary, the $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ resulting after stratification to $S = 1$ can be bounded in the following ways, where $\lambda = \left[\frac{[1-NSMI(D;[W,X])][1-NSMI(S;[D,W,X])]}{[1-NSMI(D;S)][1-NSMI([W,X];S)]}\right]$ and $\Lambda = \left[\frac{[1-NSMI(D;W|X)][1-NSMI(S;[D,W]|X)]}{[1-NSMI(D;S|X)][1-NSMI(W;S|X)]}\right]$.*

1. $R^2_{D\sim W|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - \lambda^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right)$
2. $R^2_{D\sim W|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - [1 - NSMI(D;X|S=1)]\Lambda^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right)$
3. $\eta^2_{D\sim W|X,S=1} \leq \frac{1}{1-\eta^2_{D\sim X|S=1}} \times \left(1 - \lambda^{\frac{1}{p(S=1)}} - \eta^2_{D\sim X|S=1}\right)$
4. $\eta^2_{D\sim W|X,S=1} \leq \frac{1}{1-\eta^2_{D\sim X|S=1}} \times \left(1 - [1 - NSMI(D;X|S=1)]\Lambda^{\frac{1}{p(S=1)}} - \eta^2_{D\sim X|S=1}\right)$

where $R^2_{D\sim X|S=1}$ or $\eta^2_{D\sim X|S=1}$ is estimated from the data. We can approximate or inform the choice of $NSMI(D;X|S=1)$ using the estimated $R^2_{D\sim X|S=1}$, $\eta^2_{D\sim X|S=1}$, or the related L-measure.[18] *These bounds are all analogous to bound 1 in Theorem 2. Analogs to bounds 2 - 6 in Theorem 2 could also be formed.*

Normalized scaled mutual information (NSMI) Theorems 2 and 3 ask us to reason about NSMI values. So how should we think about NSMI? NSMI is a mutual information based measure of dependence between random variables. Therefore, it measures the full dependence relationship of two random variables, not just the linear dependence or dependence

[18] We cannot directly estimate NSMI($D; X|S = 1$), since we cannot estimate $\mathbf{\Omega}$. See Appendix B Section B.6 for discussion of $\mathbf{\Omega}$.

related through the conditional expectation function. We show in Appendix B Section B.6 that, for two random variables (X, Y), *NSMI*$(X;Y)$ *can be thought of as a measure of the* ***proportion*** *of the* ***certainty*** *in the outcomes of* X*, after we learn the value of* Y*, that is* ***gained*** *as a result of learning the value of* Y*.* (As opposed to the proportion of the certainty in the outcomes of X, after we learn the value of Y, that existed before we learned the value of Y.) This echoes the typical interpretation of mutual information as the "amount of information" obtained or gained about X as a result of learning the value of Y. NSMI can indeed also be interpreted just as a normalized and scaled version of mutual information; but it also has this additional interpretation that is similar to an R^2 being the proportion of variance in one variable explained by another variable. NSMI, and the L-measure it is based on, is a useful measures of dependence between random variables in that it satisfies the following properties discussed in Rényi (1959), Smith (2015), Lu (2011), and others as the properties possessed by "an appropriate measure of dependence."[19][20]

1. NSMI is defined for arbitrary pairs of random variables.[21]
2. NSMI is symmetric.
3. NSMI takes values between 0 and 1.
4. NSMI equals 0 if and only if the variables are independent.

[19] Mutual information satisfies properties 1, 2, 4, and 6. Squared Pearson correlation (i.e., R^2) satisfies properties 1, 2, 3, 5, and 7. η^2 also does not satisfy all of these properties. See Rényi (1959) for further discussion.

[20] The transformation $\ell^2(\text{MI}(X;Y)) = 1 - \exp(-2 \times \text{MI}(X;Y))$ ensures that properties 2, 3, 6, and 7 are satisfied; it is the transformation that turns mutual information into an R^2 for Gaussian distributed variables. The transformation $L^2(\text{MI}(X;Y)) = 1 - \exp(-2 \times \text{IF} \times \text{MI}(X;Y))$ is the square of Lu (2011)'s L-measure, where IF is chosen to ensure that properties 1 and 5 are satisfied, while also maintaining properties 2, 3, 6, and 7. The transformation $\text{NSMI}(X;Y) \triangleq L^2_{\boldsymbol{\Omega}}(\text{MI}(X;Y)) = 1 - \exp(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(X;Y))$ is our normalized and scaled measure of mutual information, where $\boldsymbol{\Omega} \geq 0$ is also chosen to ensure that $\text{NSMI}(D; \mathbb{E}[D|W, S = 1]|S = 1) = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} = \eta^2_{D\sim W|S=1}$ is satisfied, while also maintaining properties 1 through 7. Lu (2011) demonstrates that properties 1 through 7 hold for the L-measure. Given this, it is trivial to see that they also hold for NSMI. See Appendix B Section B.6 for discussion of $\boldsymbol{\Omega}$ and IF.

[21] When there are multiple unobserved variables contained in W, we can consider them in combination and consider their NSMI. That is, let $W = \{W_1, W_2, \ldots, W_k\}$ and consider NSMI like $\text{NSMI}(S; [D, W]) = \text{NSMI}(S; [D, W_1, W_2, \ldots, W_k])$ or $\text{NSMI}(D; W) = \text{NSMI}(D; [W_1, W_2, \ldots, W_k])$.

5. NSMI equals 1 if and only if the variables have a strict dependence.
6. NSMI is invariant to marginal, one-to-one transformations of the variables.
7. If the variables are Gaussian distributed, then NSMI equals their R^2.

Laarne et al. (2021) notes that "MI is invariant under monotonic transformations of variables. This means that the MI correlation coefficient of a non-linear model (X, Y) matches the Pearson correlation of the linearized model $(f(X), g(Y))$. General conditions for f and g are described in" Ihara (1993). The "MI correlation coefficient" discussed in Laarne et al. (2021) is defined in a similar way to NSMI for continuous variables. Thus, NSMI might be thought of as the R^2 for a linearized model.[22]

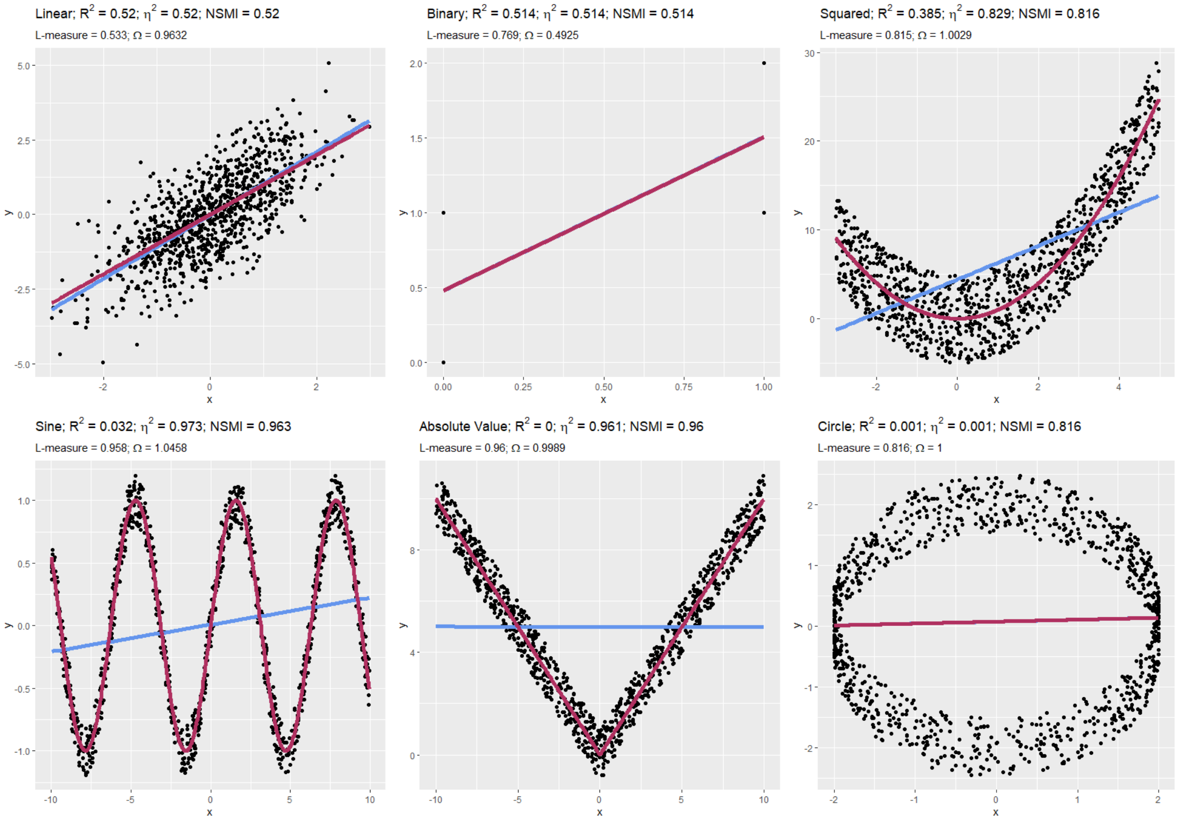
We also provide examples to help readers gain some familiarity with NSMI. In Figures 3.3 and 3.4, we show 12 different types of bivariate relationships with the corresponding R^2, η^2, and NSMI. In these examples, we estimate NSMI using the `rmi` and `infotheo` `R` packages and η^2 with the `KRLS` `R` package using simulated samples of 1000 data points. (Michaud, 2018; Meyer, 2014; Hainmueller and Hazlett, 2014; Ferwerda et al., 2017) There is estimation error in these, since mutual information can be difficult to estimate in practice, but the figures should still be informative.[23] We see that NSMI is larger that η^2 but is often very comparable. When η^2 does a poor job of capturing the full relationship between the variables, NSMI can be much larger than η^2. Lu (2011)'s L-measure is close to or larger than NSMI. So it is possible to reason about the L-measure as an approximation or as a bound on NSMI.

[22] It is worth noting that, although we might be more comfortable thinking about correlations and R^2's, they are not necessarily capturing what we expect. First, correlation and R^2 capture only the strength of linear association; these do not necessarily capture an intuitive sense of dependence but one restricted to linear relationships. Also, "Mutual Information is a nonlinear function of ρ which in fact makes it additive. Intuitively, in the Gaussian case, ρ should never be interpreted linearly: a ρ of $\frac{1}{2}$ carries $\approx$ 4.5 times the information of a $\rho = \frac{1}{4}$, and a ρ of $\frac{3}{4}$ 12.8 times!" (Taleb, 2019) "One needs to translate ρ into information. See how $\rho = 0.5$ is much closer to $[\rho =]0$ than to a $\rho = 1$. There are considerable differences between .9 and .99." (Taleb, 2019) See Figure B.1 for a series of plots that illustrate how changes in correlation and R^2 compare to changes in mutual information for standard Gaussian random variables. See Figure B.2 for a plot of the relationship between mutual information and R^2 for Gaussian variables, this is also the normalization curve we use. Mutual information can capture our intuitive sense of dependence better than correlation and R^2 even in the simple Gaussian case.

[23] In addition, we present the L-measure and $\boldsymbol{\Omega}$. See Appendix B Section B.6.

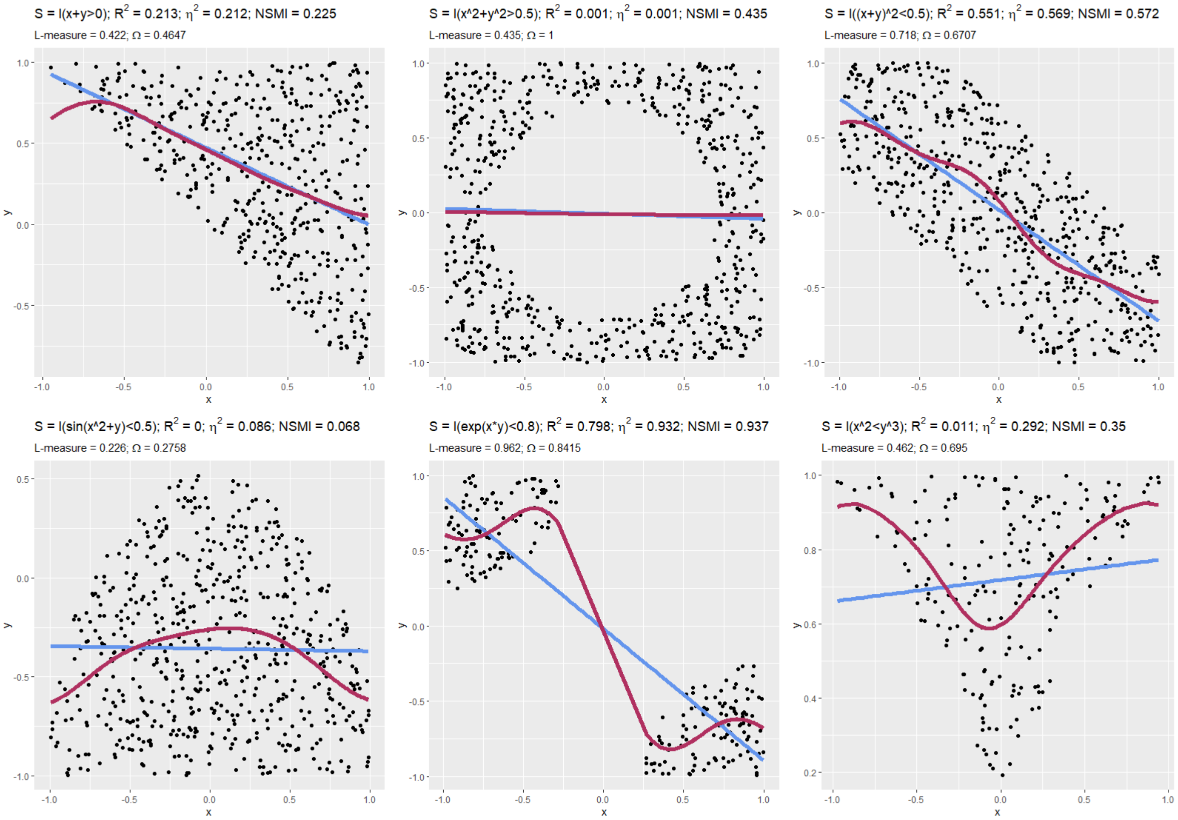
See Appendix B Section B.6 for more detail on NSMI and the L-measure. See Figure B.1 for a series of plots that illustrate how changes in correlation and R^2 compare to changes in mutual information for standard Gaussian random variables. In the Gaussian case, NSMI equals R^2; and so interpretation of NSMI should be familiar.

Figure 3.3: NSMI Examples. These are generated with various linear and non-linear relationships between x and y. The blue line is an linear fit. The red line is a flexible fit or the true non-linear relationship.



Discussion of bounds Theorems 2 and 3 contain several bounds. All the bounds presented in Theorem 3 correspond to bound 1 from Theorem 2. Analogs to bounds 2-6 from Theorem 2 can also be created for the case where there are covariates X. We expect that the simplest bounds to use will often be bounds 4-6 from Theorem 2. See the worked example below for an example of how bound 5 from Theorem 2 can be adapted to include covariates.

Figure 3.4: More NSMI Examples. These are generated by selecting a non-random sample from two uniform random variables. S is the sample selection variable. The blue line is an linear fit. The red line is a flexible fit.



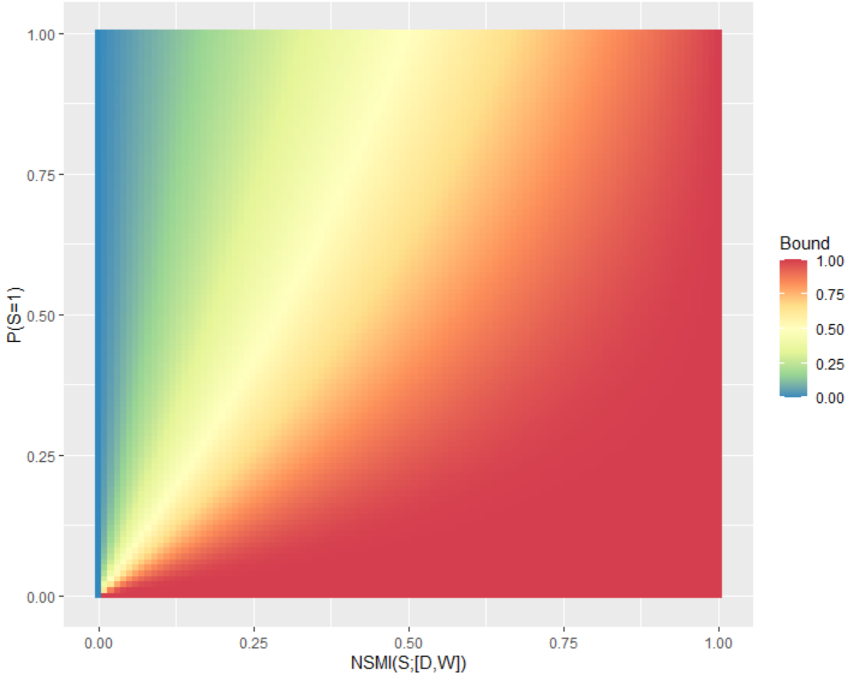
Bounds 1 through 3 in Theorem 2 are tighter than bounds 4 through 6, but require additional sensitivity parameters as well as some knowledge about how mutual information works. That is, since some of the NSMI quantities are related in the bounds in Theorem 2, users need to take care to reason about coherent combinations of the NSMI quantities. In particular, the bounds all take the form $1-(\tau)^{\frac{1}{p(S=1)}}$ but with different τ; τ must take a value between 0 and 1. This reflects the fact that $1-(1-\text{NSMI}(W;D|S))^{\frac{1}{p(S=1)}}$ equals bounds 1 through 3 and $\text{NSMI}(W;D|S)$ takes values between 0 and 1. This, in turn, reflects that $\text{MI}(D;W|S) = \text{MI}(D;W) + \text{MI}(S;[D,W]) - \text{MI}(S;D) - \text{MI}(S;W) \geq 0$. For this reason, we encourage users unfamiliar with mutual information to use bounds 4 through 6, where the condition that $\tau \in [0,1]$ will always be satisfied given NSMI values between 0 and 1. Bounds

4 and 5 are tighter than bound 6. If W and D are assumed to be marginally independent, then $\text{NSMI}(D; W) = 0$. Which bound is most useful depends on the relationships that practitioners feel most comfortable reasoning about in terms of NSMI's.

We consider in detail bound 6 from Theorem 2. This bound is an expression of normalized scaled mutual information for the marginal mutual information between D and W, for the mutual information between S and $[D, W]$ together, and the probability of selection, $P(S = 1)$.[24] As we saw in the case of binary random variables and truncated normal random variables, we have an expression in terms of structural (i.e., causal) relationships between the variables in the full population. In Figure 3.5, we show how bound 6 from Theorem 2 changes for different values of $\text{NSMI}(S; [D, W])$ and $p(S = 1)$. For this, we assume that W, D are marginally independent and so $\text{NSMI}(D; W) = 0$ and the bound becomes $B \triangleq 1 - (1 - \text{NSMI}(S; [D, W]))^{\frac{1}{p(S=1)}}$. As $p(S = 1) \to 1$, $B \to \text{NSMI}(S; [D, W])$. As $p(S = 1) \to 0$, $B \to 1$. As $\text{NSMI}(S; [D, W]) \to 1$, $B \to 1$. As $\text{NSMI}(S; [D, W]) \to 0$, $B \to 0$. These dynamics are easy to see in the expression for the bound itself. They reflect the bounds on $\text{MI}(W; D|S = 1)$ from Theorem 1 that we then scale and normalize. It is worth noting that small probabilities of selection can lead to bounds close to 1, regardless of the value for $\text{NSMI}(S; [D, W])$. $\text{MI}(D; W|S) = p(S = 1)\text{MI}(D; W|S = 1) + p(S = 0)\text{MI}(D; W|S = 0)$ will be dominated by $\text{MI}(D; W|S = 0)$ when $p(S = 1)$ is small. Further, we use $\frac{\text{MI}(D;W|S)}{p(S=1)}$ to bound $\text{MI}(D; W|S = 1)$; a small $p(S = 1)$ ca lead to large values for $\frac{\text{MI}(D;W|S)}{p(S=1)}$, even when $\text{MI}(D; W|S)$ is small.

[24] $P(S = 1)$ can be thought of as the proportion of the population captured by the sub population for which our selected sample is a representative. It is not necessarily the size of our data sample relative to the size of the population. Note that it is important to have a clear sense of the population from which the sample has been selected here, but this is already required to be able to think about the sample selection mechanism in the first place and hence know whether conditioning on W will yield conditional ignorability or not.

Figure 3.5: Bound 6 from Theorem 2 on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ given values for NSMI$(S;[D,W])$ and $p(S=1)$ and assuming NSMI$(D;W)=0$



3.3 Worked example

We now turn our attention to an application of the proposed methods. Hazlett (2020) considers the effect of being directly harmed in the conflict in Darfur in early 2000s on attitudes about peace using a survey of individuals in refugee camps. Motivated by "long-standing neglect of the region by the central government" and a "history of attacks on civilians by both the Sudanese army and irregular militia," rebel groups attacked government air force base in February 2003. The government responded with an operation aiming to "punish, kill, or displace" ethnic groups thought to be supporting the rebellion. The article argues that "violence was targeted by village and gender but was indiscriminate beyond this" and that the "evidence is consistent not with the 'angry' response but rather with claims of a 'pro-peace' or 'weary' effect of exposure to violence." The paper also describes that "Most refugees or internally displaced persons left their homes during 2003 to 2004. A large number of those in the Western regions of West Darfur made the decision to cross the border into eastern Chad. Very few of these refugees had returned home by the time of this survey in mid-2009, when approximately 250,000 Darfurians were registered in refugee camps in eastern Chad." The study relies on data "drawn from a survey conducted between April and June of 2009 by the 'Darfurian Voices' team with support of the US Department of State... The full survey was thus representative of adult refugees (eighteen years or older) from Darfur, living in the twelve Darfurian refugee camps in eastern Chad at the time of sampling." Let us assume the population is the set of refugees who made it to the camps at some point. Since some individuals left these camps before the time of the survey, we have a threat of bias from sample selection.

The study controls for things like village, gender, and other important covariates in estimating the causal effect of being directly harmed on attitudes about peace. While adjustment for these covariates is important, we still need to consider sample selection. Being harmed may effect whether someone re-entered the conflict (and hence was not captured in the survey). An individual's pro-peace predisposition (before the conflict, an unobserved variable) may be a common cause of both whether they re-entered the conflict and their peace attitudes at the time of the survey. This may create a non-causal path running from harm to pro-peace predisposition to attitude about peace that could threaten the internal validity of estimated effects. In Figure 3.6, D is direct harm, Y is attitudes about peace, W is pro-peace predisposition (before the conflict), S is being in the survey from the refugee camp (i.e., did not re-enter the conflict), and X is observed covariates like village and gender. Hazlett (2020) is able to adjust for the observed covariates, but the path $D \dashrightarrow Z \rightarrow Y$ cannot be blocked since pro-peace predisposition is not observed. That is, we are able to estimate an effect controlling for age, gender, village, and other covariates. But this effect could be biased by the spurious relationship created by sample selection as a collider between direct harm (D) and pro-peace predisposition (W). Using OLS regression (among other estimation strategies), Hazlett (2020) estimates that peace index is .09 to .10 units higher among those directly harmed. See Table 3.1.

Figure 3.6: Possible sample selection bias in Hazlett (2020)

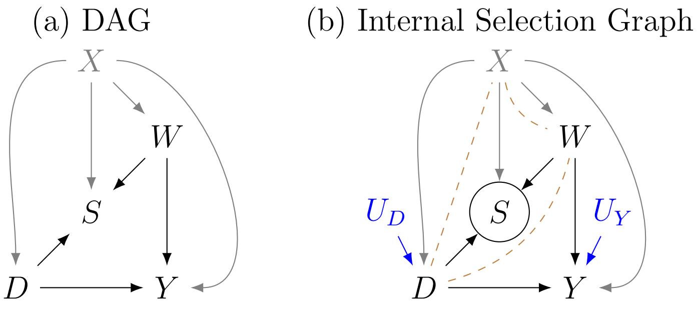


Table 3.1: Causal effect estimate from Hazlett (2020)

Outcome: *peace factor*

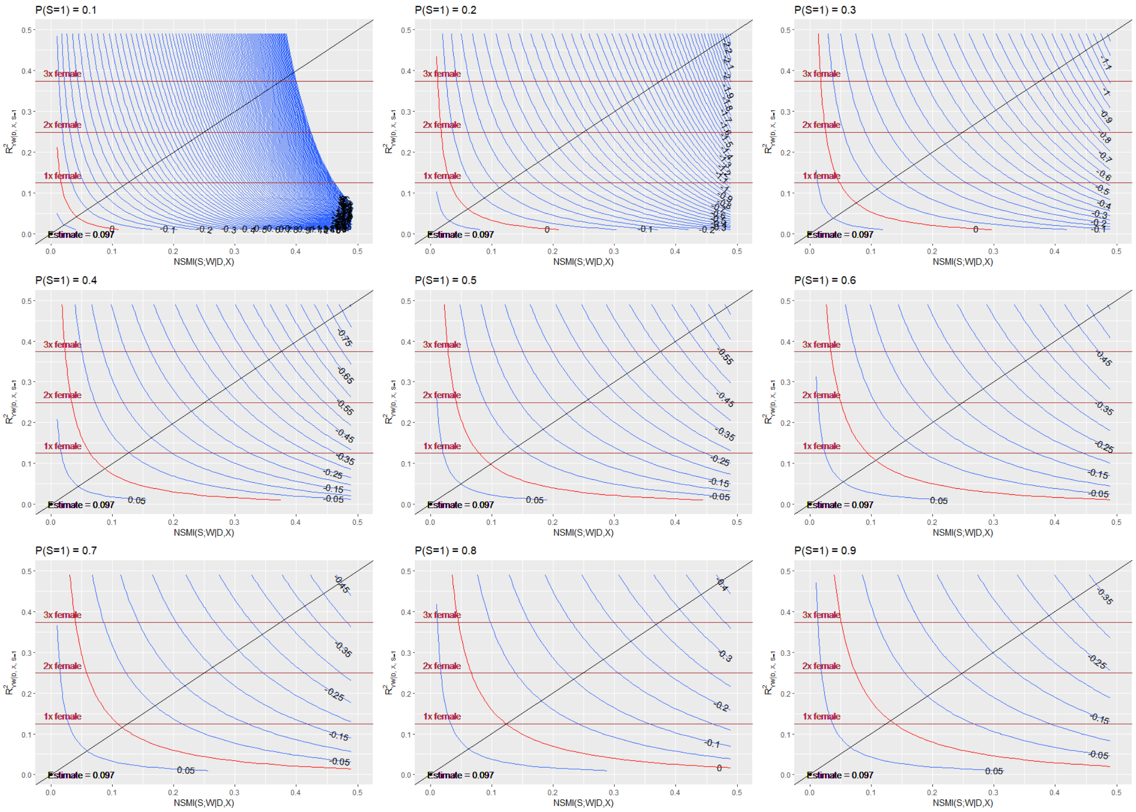
Treatment:	Est.	S.E.	t-value
directly harmed	0.097	0.023	4.184

df = 783

The process that would drive individuals back into the conflict would "act more powerfully for men of fighting age because in this context, few women or elderly participate directly in the armed opposition groups. If such a process drove the results, we would see the apparent effect most strongly among young men but should see little or no apparent effect among women or the elderly who are far less likely to join the opposition. This is not the case." (Hazlett, 2020) We might then claim that the effect of direct harm on peace attitudes among women and the elderly is perhaps not biased by this sample selection mechanism. But this sample selection mechanism might not allow us to obtain an internally valid effect estimate for fighting age men or the full set of refugees captured in the survey.

We can use the bounds from Theorems 2 and 3 in a sensitivity analysis for linear regression following Cinelli and Hazlett (2020) using the software from Cinelli et al. (2020). This omitted variable bias based sensitivity analysis requires that we consider hypothetical values for $R^2_{W\sim D|X,S=1}$ and $R^2_{Y\sim W|D,X,S=1}$. Inspecting Figure 3.6, we see that $R^2_{Y\sim W|D,X,S=1}$ captures just the causal path $W \to Y$, that is, the strength of the relationship between pro-peace predisposition (W) and attitudes about peace after the conflict (Y) after controlling for observed covariates (X) like village and gender. This is a structural relationship that we will be able to build some intuition about. On the other hand, reasoning about $R^2_{W\sim D|X,S=1}$ is more difficult. This captures the path $W \cdots D$ and relates to the strength of the relationship between pro-peace predisposition (W) and direct harm (D) within the selected sample of refugees that did not reenter the fight after controlling for observed covariates (X) like village and gender. These variables do not have a direct relationship in the population that we can build intuition about that would allow us to directly reason about $R^2_{W\sim D|X,S=1}$. In fact, the assumption in Figure 3.6 is that pro-peace predisposition (W) is independent of direct-harm (D) in the population, conditional on the observed covariates like village and gender. So their entire relationship is created as a result of conditioning on a collider due to sample selection.

Figure 3.7: Sensitivity analysis contour plots for Darfur example. Contours represent revised effect estimates.



Given the difficulty in interpreting and building intuition for $R^2_{D\sim W|X,S=1}$, we use bound 2 from Theorem 3 to bound $R^2_{D\sim W|X,S=1}$, where we assume that D and W are independent conditional on X in the population. We also choose a bound analogous to bound 5 from Theorem 2. The bound we use is therefore

$$\begin{aligned} R^2_{W\sim D|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \\ \left(1 - [1-\mathrm{NSMI}(D;X|S=1)]\,[1-\mathrm{NSMI}(S;W|D,X)]^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right), \end{aligned} \tag{3.10}$$

where we estimate $R^2_{D\sim X|S=1}$ from the data and approximate $\mathrm{NSMI}(D;X|S=1)$ based on that estimate. We simply assume that we have the worst case where $R^2_{W\sim D|X,S=1}$ equals this bound and substitute the bound into the omitted variable bias expression provided in Cinelli and Hazlett (2020) and use this to calculate revised estimates given hypothesized values of $R^2_{YW|D,X,S=1}$, $\mathrm{NSMI}(S;W|D,X)$, and $p(S=1)$. Contour plots that show the surface of revised estimates for the breadth of values these three sensitivity parameters can take are displayed in Figure 3.7.

We will make some assumptions on $p(S=1)$ and on $R^2_{Y\sim W|D,X,S=1}$ for the purpose of illustration so that we may focus on the remaining parameter, $\mathrm{NSMI}(S;W|D,X)$. We emphasize these assumptions are for illustration only; subject matter experts may debate these. We can take $1-p(S=1)$ to represent the portion of refugees that reentered the fighting and were, therefore, not eligible to be captured by the survey. We might suppose that no more than 20% of individuals reentered the fighting. If we believe this is a plausible bound on $1-p(S=1)$, we might then consider the contour plot at the bottom middle of Figure 3.7. Suppose also that we believe that no unobserved variable, including pro-peace predisposition, will explain more of the outcome than the female variable. We show 1x, 2x, and 3x how much the female variable explains of the outcome in the contour plots. In the $p(S=1)=0.8$ panel of Figure 3.7, we see that assuming that $R^2_{Y\sim W|D,X,S=1}$ equals the partial R^2 between the female variable and the outcome, peace index, an $\text{NSMI}(S; W|D, X)$ of about 0.13 or so would bring our effect estimate to zero.

Do we think that an $\text{NSMI}(S; W|D, X)$ of 0.13 or more is plausible? $\text{NSMI}(S; W|D, X)$ can be thought of as the proportion of the certainty inherent to whether someone re-enters the fight, given that we know whether they were directly harmed and their pro-peace predisposition (as well as their gender, village, and other observed covariates), that is gained as a result of learning their pro-peace pre-disposition. If the variables shared a joint Gaussian distribution, then this would correspond to an R^2 of 0.13. The question becomes: to what extent does reentering the fight depend on pro-peace predisposition, after controlling for the observed covariates and direct harm? How much of the information we have about reentering the fight is gained as a result of learning the value of this variable? The key is that this is a substantive question that subject matter experts can debate. We no longer are in a situation in which the threat of bias from sample selection means that we cannot gain anything from our effect estimates. Instead, we can appeal to external knowledge about a substantive question. If we believe that pro-peace predisposition does not have a strong relationship with whether or not someone was captured in the survey, then perhaps an $\text{NSMI}(S; W|D, X)$ of 0.13 or less is plausible. This type of analysis shifts criticism away from whether or not there exists a threat of sample selection bias towards discussions that attempt to discern whether some substantive structural relationships meet a threshold level of strength that would change the conclusions of the estimated effect. We hope that this example provides some guidance to how such sensitivity analysis can be conducted in practice.

3.4 Discussion

We have proposed an approach to generalizing the omitted variable bias sensitivity analyses of Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) for evaluating the threats posed to the internal validity of causal effects by sample selection. We saw how the non-parametric bounds on sensitivity parameters require users to reason about structural (causal) relationships comprising the sample selection mechanism in the population, mirroring expressions we derived for the sensitivity parameters in simple parametric settings. We also worked through an application illustrating how the approach might be used. Other approaches to sensitivity analysis for sample selection also exist. Smith and VanderWeele (2019) discuss approaches with sensitivity parameters that capture the relationships between unobserved variables and the observed variables, as we do here. However, building intuition for their parameters may not be as familiar as using R^2s or η^2s. Moreover, in their discussion of "the selected population as the target population," they only provide heuristic guidance on how researchers might deal with the counterintuitive nature of sensitivity parameters capturing associations between marginally independent variables that are made dependent due to sample selection. Thompson and Arah (2014) also present an approach for sensitivity analysis for sample selection. This approach requires specifying sensitivity parameters that filter into a model of the selection mechanism. Greenland (2003); Hernán et al. (2004); Elwert and Winship (2014); Infante-Rivard and Cusson (2018); Arah (2019), and many others also provide insightful discussions into sample selection bias and potential remedies. The benefit of our approach is the improved interpretation of the sensitivity parameters and connections to the very useful, existing omitted variable bias based sensitivity analysis frameworks of Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) for which software already exists.

CHAPTER 4

Instrumental variables and sample selection

When researchers are interested in the causal effect of a treatment on an outcome but are not confident that conditioning on observed covariates eliminates all confounding of the treatment-outcome relationship, they might appeal to an instrumental variables identification strategy. The instrumental variables ("IV") identification strategy attempts to leverage variation in a variable that is associated with the treatment but not directly with the outcome (this variable is called the instrument) to try to understand the causal relationship between the treatment and the outcome. (Imbens and Angrist, 1994; Angrist et al., 1996; Hernán and Robins, 2006; Pearl, 2001, 2009; Baiocchi et al., 2014; Hernán and Robins, 2020) In it's simplest form, this boils down to looking at how the outcome and the instrument are associated and how the treatment and the instrument are associated and then trying to use these components to get at how the treatment and outcome are causally related. To go from associations between the instrument and outcome and the instrument and treatment to a causal relationship between the treatment and outcome requires restrictions on the causal relationships between the three variables. The specific restrictions are discussed in detail in subsequent sections but at a minimum these include "ignorability" and "relevance" to bound the treatment effect. (Pearl, 2009; Balke and Pearl, 1994a) Other assumptions can allow for point identification. These restrictions are often demanding and do not hold in many cases. But even when these restrictions do appear to be met, we must also consider how the sample of units available to study was selected from possible larger populations and whether this sample selection mechanism could bias the effect estimate. In particular, we concern ourselves with whether or not an estimated effect obtained from a sample drawn in some non-random way is an unbiased estimate of the causal effect averaged over the members of the selected group, which is traditionally referred to as “internal validity.” (Campbell, 1957; Campbell and Stanley, 1966; Cook and Campbell, 1979; Shadish et al., 2002)

Many other authors have pointed out that sample selection can violate the assumptions of instrumental variables approaches. Canan et al. (2017); Swanson et al. (2015); Swanson (2019); Hughes et al. (2019); Ertefaie et al. (2016); Gkatzionis and Burgess (2018); Hernán and Robins (2020); Elwert and Segarra (2022) all discuss sample selection and instrumental variables. Canan et al. (2017) discusses how one form of sample selection can violate IV assumptions. Hughes et al. (2019) provides a broader view into how sample selection can violate IV assumptions. They provide several examples, run simulation studies, and provide some guidance and description on the reason violations arise in their examples. Swanson (2019) discusses broad questions about how sample selection can bias IV studies and provides guidance for applied researchers. They discuss threats posed by sample selection problems that are unique to IV (e.g., selection on the treatment can be a problem for IV whereas it is not typically for a simple covariate adjustment approach). They also mention that not all types of sample selection that might threaten internal validity in the context of other designs threaten internal validity for instrumental variables. Hernán and Robins (2020) in their discussion of instrumental variables mention that sample selection can violate instrumental variables assumptions and also briefly mention some interesting cases that we analyze further in this paper. There are various papers that focus on applications that explore specific examples. Of particular interest is where researchers are interested in comparing two treatment levels and select only units that receive either of these treatment levels, but more than two treatment levels exist (Swanson et al., 2015). Sheehan et al. (2008) provides good examples of proxy and confounded instruments, as does Hernán and Robins (2020). Swanson and Hernán (2013) suggest a checklist for reporting IV conditions and results. These authors, however, do not provide systematic guidance for if/when instrumental variables can apply in an arbitrary causal graph with any sample selection mechanism. Though other authors have provided comprehensive guidance for instrumental variables. Van Der Zander et al. (2015); Van Der Zander and Liśkiewicz (2016); Kumor et al. (2020) present graphical approaches to finding instruments and their generalizations, but do not focus on how sample selection relates to these. Galles and Pearl (1998); Pearl (2001, 2009); Elwert and Segarra (2022) discuss conditional instruments and graphical criteria for them. While Pearl (2001); Elwert and Segarra (2022) focus primarily on linear models, Galles and Pearl (1998); Pearl (2009) discuss potential outcomes. Further, Elwert and Segarra (2022) focus on sample selection resulting from conditioning on a descendant of the treatment; we extend our analysis beyond such cases and provide clear graphical guidance on when instrumental variables can be used for any sample selection mechanism. Elwert and Segarra (2022) provide exact expressions for sample selection bias in linear models under a few important sample selection mechanisms.

We aim to bridge the gap between these two sets of papers. We use the internal selection graph from Chapter 2 to simply and intuitively visualize the effects that sample selection has on the relationships between variables in the causal graph. We restate the graphical criterion from Galles and Pearl (1998); Pearl (2001, 2009) highlighting the special role that sample selection plays. We therefore provide a comprehensive guide to how sample selection can threaten the internal validity of instrumental variables approaches. This generalizes the discussions of sample selection as a potential source of bias for instrumental variables found in Canan et al. (2017); Hughes et al. (2019); Swanson (2019); Elwert and Segarra (2022) and elsewhere. We then use these tools to we explore many interesting implications of sample selection for instrumental variables evaluating both threats and opportunities.

4.1 Instruments under sample selection

We are interested in the causal effect of a treatment, D, on an outcome, Y, *for units in the selected sample*. We will use a binary variable, S, to denote sample selection. Our approach is grounded in structural causal models (SCM; Pearl (2009)), potential outcomes

(Splawa-Neyman et al. (1990), Rubin, 1974, 1978, 1990), and directed acyclic graphs (DAGs; Pearl (2009)). Potential outcomes are solutions to the equations in SCMs, under intervention. The equations and variables in SCMs correspond to the edges and nodes in DAGs. A potential outcome, Y_{di}, is the value that the variable Y would have taken for unit i, if the variable D for unit i had been set, possibly counterfactually, to the value d. The unit-level causal effect of setting D to d relative to D to d' is $\tau_i = Y_{di} - Y_{d'i}$. The fundamental problem of causal inference, however, is that we are never able to observe more than one of the potential outcomes for a given unit and so cannot calculate unit level causal effects. (Rubin, 1978; Holland, 1986; Imbens and Rubin, 2015; Westreich et al., 2015) Our target will therefore instead be the average of τ_i taken over some subset of the units in our sample, in which some units have $D = d$ and others have $D = d'$. An estimation strategy is said to be "internally valid" if it can unbiasedly or consistently estimate such quantities. See Chapter 2 for additional discussion of these concepts.

Instrumental variables can be used to bound causal effects of the treatment on the outcome in the presence of relationships between the treatment and outcome other than the causal relationship. This can be a powerful capacity, as ruling out unobserved confounding and other forms of non-causal relationships between the treatment and outcome can be difficult. The key to an instrumental variables approach is the presence of a variable (that we call an instrument, IV) that is associated with the treatment, D, but that is not otherwise associated with the outcome, Y. We want the instrument to covary with the treatment but only covary with the outcome as a result of the causal relationship between the treatment and the outcome. When such a variable exists, we are able to use the association between the instrument and the outcome as well as the association between the instrument and the treatment to bound (or, with additional assumptions, identify) a causal effect of the treatment on the outcome.

We alter the definition of an instrument found in Pearl (2009) to explicitly state that we must restrict ourselves to the selected sample, that is, we must condition on $S = 1$. The following definition is adapted from Pearl (2009), Definition 7.4.1.[1]

Definition 5 (Instruments, Relevance, and Ignorability)**.** A variable IV is an instrument relative to the total effect of D on Y within the stratum $S = 1$ if there exists an X, unaffected by D, such that the following hold.

1. (relevance) $D \not\perp\!\!\!\perp IV | X, S = 1$
2. (ignorability) $Y_d \perp\!\!\!\perp IV | X, S = 1$

Relevance captures the idea that, in order to study the relationship between the treatment and outcome, the instrument must be associated with the treatment. Otherwise, we cannot use the instrument to isolate any of the variation between the treatment and the outcome. Ignorability captures the idea that, while we want the instrument to be associated with the treatment, we do not want it to directly cause the outcome or to be related with the outcome other than through the causal relationship between the treatment and the outcome. If it were associated with the outcome in one of these ways, then we could not isolate the causal relationship between the treatment and the outcome from the association between the instrument and the outcome. Our alteration to Pearl's definition simply makes explicit that we want these conditions to hold in the selected sample.

How do we know whether relevance and ignorability hold? In practice, we can never be certain that some set of covariates will provide the relevance and ignorability we need. The

[1] This definition is by no means the only way to define an instrument. Pearl (2009) also provides conditions that are purely graphical and do not involve potential outcomes. Angrist et al. (1996); Lousdal (2018) require that the IV - D relationship is causal and unconfounded, which Pearl (2009) points out is unnecessary in general. Hernán and Robins (2006); Hernán and Robins (2020) split ignorability into two conditions, one of which is the well known ”exclusion” restriction: $Y_{d,iv} = Y_{d,iv'} = Y_d$ and $Y_{iv,d} \perp\!\!\!\perp IV$. As shown in the main text, we can combine exclusion and ignorability into $Y_d \perp\!\!\!\perp IV$; see Hernán and Robins (2020). (Proof: $Y_{d,iv'} = Y_d$ and $Y_{iv,d} \perp\!\!\!\perp IV \implies Y_d \perp\!\!\!\perp IV$: $(Y_{iv,d} = Y_d) \perp\!\!\!\perp IV$. $Y_d \perp\!\!\!\perp IV \implies Y_{d,iv'} = Y_d$ and $Y_{iv,d} \perp\!\!\!\perp IV$: Suppose $Y_{iv,d} \neq Y_d$. Then there is a path from IV to Y that does not run through D. This means that IV is a common cause of Y and D and so $Y_d \not\perp\!\!\!\perp IV$, a contradiction. So $Y_{iv,d} = Y_d$, which in turn means $(Y_{iv,d} = Y_d) \perp\!\!\!\perp IV$.) Greenland (2000); Didelez and Sheehan (2007a,b); Sheehan et al. (2008) use somewhat different conditions that invoke an unobserved confounder explicitly. Swanson et al. (2018) discuss bounds that can be found for the ATE under various alternative IV conditions. We adopt what they describe as the "least restrictive" set of IV conditions here.

onus is on researchers to make plausible arguments for relevance and ignorability. To aid in this, we can build a causal graph that captures how the treatment and outcome causally relate to each other and relevant covariates. Causal graphs allow us to visualize dependencies and independencies between variables in terms of a path separation criterion, *d-separation*. (Pearl, 2009) Two sets of nodes, D, Y, in a graph G are said to be *d-separated* by a third set, Z, if every path from any node $D_0 \in D$ to any node in $Y_0 \in Y$ is blocked. A path is blocked by Z if either [1] some W is a collider[2] on the path between D, Y and $W \notin Z$ and the descendants of W are not in Z or [2] W is not a collider on the path but $W \in Z$. See Pearl (2009) for further introduction to causal graphs and structural causal models. Simple rules can then be used to determine when relevance and ignorability hold.

It is acceptable for IV and D to be associated due to a common cause or some other relationship, as long as all the association between IV and Y is through the causal path from D to Y. IV and D need only be associated conditional on X and $S = 1$, we do not require that IV directly causes D. However, we might consider a few sub-types of instruments. "Causal" instruments are those for which there is an unconfounded causal path $IV \rightarrow D$. "Proxy" instruments are those for which the association between IV and D flows through a path like $IV \leftarrow U^* \rightarrow D$, where U^* is a causal instrument and IV is a proxy for U^*. Van Der Zander et al. (2015) define an "ancestral" instrument as one for which conditioning on a variable creates the relevance needed for an instrument. This might look like $IV \rightarrow X \leftarrow U^* \rightarrow D$, where, again, U^* is a causal instrument and IV is a proxy for U^*, conditional on X.

[2] A collider is a node in the graph into which two arrows point: $A \rightarrow S \leftarrow B$. See Pearl (2009) for an introduction to causal graphical models and colliders. Conditioning on a collider or a descendant of a collider can induce an association between the parents of the collider. Shahar and Shahar (2017) discuss the conditions under which such an association is created. Since our approach is non-parametric and graphical, we assume such an association is created when sample selection is a collider or a descendant of a collider.

4.2 Proposal

We propose using extended causal graphs that explicitly show how sample selection alters the relationships between variables in the sample. We also provide rules (graphical criteria) for using these graphs to determine when relevance and ignorability hold. In doing so, we revise the existing graphical approaches for instruments to highlight the special role of sample selection. At the same time, we formalize and generalize recent discussions of sample selection in the instrumental variables context.

4.2.1 Internal selection graphs

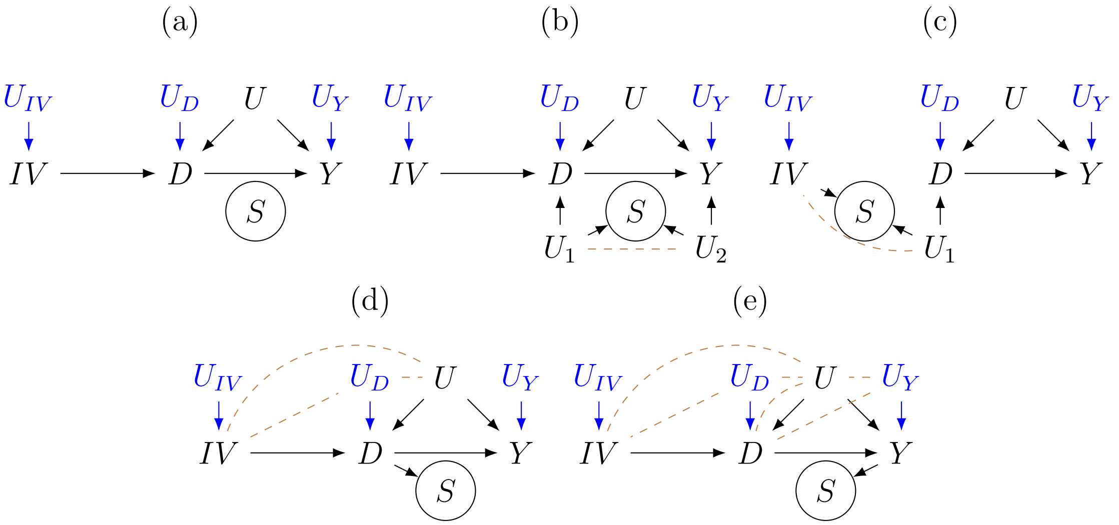
The key to our proposal is to graphically represent the ways in which sample selection alters the relationships in the selected sample. We do this by slightly altering the internal selection graphs defined in Chapter 2, which visually extend traditional causal graphs to represent all the ways that sample selection can change relationships between variables.

Definition 6 (Internal Selection Graph, G_S^+)**.** Let G be the DAG induced by a SCM.

1. Create G_S by adding an appropriately connected binary selection node, S.
2. Draw a circle around S to clearly indicate that we must limit our analysis to $S = 1$.
3. Add to G_S any node which is a parent of the treatment or a parent of a descendant of the treatment. Add to G_S any node which is a parent of the potential instrument or a parent of a descendant of the potential instrument. (U_S, the background factors contributing to selection, can be excluded.)
4. Add a dashed undirected edge between all variables between which S is a collider or an ancestor of S is a collider. We will call these dashed, undirected edges *bridges*.

Call the resulting graph an *internal selection graph, G_S^+*. (These graphs are similar to those discussed in Daniel et al. (2012) and Chapter 2.)

Figure 4.1: Examples of Internal Selection Graphs



The key features of internal selection graphs[3] are the inclusion of an encircled sample selection node, specific background variables, and bridges that capture the statistical associations that result from sample selection. These additions ensure sample selection and the changes it requires for identification are visualized in the graph and can be analyzed easily. See Figure 4.1 for examples. We will differentiate between a few types of paths. Following the discussion in Chapter 2, d-separation is defined in the same way for these paths as for regular paths, since colliders are defined in the same way. See Appendix C for details. Generalized paths are any sequence of nodes and edges (directed edges and/or bridges) where each node appears only once (e.g., $D \cdots Z \rightarrow Y$, $D \rightarrow Y$, $D \rightarrow S \leftarrow Z$, $U_D \rightarrow D \rightarrow Y$). Causal paths are any generalized path where all edges between the nodes are directed and point in the same direction (e.g., $D \rightarrow Y$, $U_D \rightarrow D \rightarrow Y$). Generalized non-causal paths are any generalized path that isn't a causal path (e.g., $D \cdots Z \rightarrow Y$). Figure 4.1(e) provides a clear example of a setting in which internal selection graphs greatly facilitate understanding how sample selection can alter relationships between the variables in the selected sample. The statistical associations created due to sample selection between many variables, as well as some of the

[3] See Chapter 2 for more discussion of internal selection graphs.

variables themselves, would be missing from the corresponding DAG.

4.2.2 Graphical criteria

Internal selection graphs already go a long ways toward facilitating analysis of instrumental variables under samples election. But how can we use internal selection graphs to determine whether relevance and ignorability hold reliably? We'll use a set of rules captured in the following graphical criteria.

4.2.2.1 Relevance

Relevance is the first condition in our definition of instruments and is perhaps the simpler of the two conditions. It captures the idea that, in order to study the relationship between the treatment and outcome using an instrument, the instrument must be associated with the treatment in some way. Otherwise, we cannot use the instrument to understand any of the variation between the treatment and the outcome. The relevance criterion is similar to condition (ii) in the graphical criterion provided in Pearl (2009) and similar to the condition (G1) in the graphical criterion provided in Elwert and Segarra (2022), but altered to indicate the special role that sample selection plays and to work with internal selection graphs.

Definition 7 (Relevance Criterion)**.** A set of nodes X and a possible instrument IV in G_S^+ satisfy the relevance criterion relative to D (treatment), and Y (outcome) if there is at least one (*causal or generalized non-causal*) path between IV and D that does not pass through S and is not blocked by X.

Theorem 4. *If a set of nodes X and a possible instrument IV in internal selection graph G_S^+ satisfy the relevance criterion relative to D (treatment), and Y (outcome), then $D \not\perp\!\!\!\perp IV|X, S = 1$.*

This result is proved in Appendix C. We need the instrument to be associated with the treatment. Whether this association manifests as a causal relationship (e.g., $IV \rightarrow D$) or a non-causal relationship (e.g., $IV \leftarrow U \rightarrow D$) is not important. For relevance to hold, we just need that the values the instrument takes are related some how to the values that the treatment takes.[4] So we just want there to be some path between these two variables. Moreover, we need this association to persist (or to arise) when we condition on both X and $S = 1$. Hence, our criterion requires that there is at least one path that does not pass through S. The paths through S become irrelevant in internal selection graphs because if S is not a collider on a path, then the path is blocked and we can ignore it; further, if S is a collider on a path, then we have already drawn a bridge between the nodes that form the collider and so we can bypass the path that actually includes S itself. If we must condition on X to achieve ignorability, we don't want it to ruin relevance. So we need a path between IV and D that is not blocked by X. Finally, there are situations in which selection and/or conditioning on X can give us relevance. Such instruments are called "ancestral" instruments, as discussed above (Van Der Zander et al., 2015). This might arise due to X being a collider in which case conditioning on X may unblock a path between IV and D; of course such a path would not be blocked by X and would satisfy the relevance criterion. No matter the specific type of path or paths that yield relevance, the key idea of relevance is simple: that IV and D should relate along some unblocked path.

4.2.2.2 Ignorability

Ignorability is the second condition in our definition of instruments and is somewhat more subtle than relevance. It captures the idea that, while we might not be very particular about how the instrument associates with the treatment, we want to be very careful about how the instrument associates with the outcome. The ignorability criterion is similar to condition (i)

[4] There are also problems associated with weak associations between the instrument and treatment. These are referred to as the weak instruments. In this paper, we focus on whether relevance holds at all and not on whether the instrument is a weak instrument. Fortunately, relevance is a condition that can actually be tested with data. Further discussion of this is also outside the scope of this paper.

in the graphical criterion provided in Pearl (2009) and similar to the conditions (G2) and (G3) in the graphical criterion provided in Elwert and Segarra (2022), but altered to indicate the special role that sample selection plays and to work with internal selection graphs.

Definition 8 (Ignorability Criterion)**.** A set of nodes X and a possible instrument IV in G_S^+ satisfy the ignorability criterion relative to D (treatment), and Y (outcome) if

1. No element of $\{X, S\}$ is a descendant of D and D is not in $\{X, S\}$.
2. X blocks every (*causal and generalized non-causal*) path between IV and Y except
   (a) those that pass through S and
   (b) those ending with a causal path from D to Y (e.g., paths between IV and Y that pass through D but where D or one of its descendants touches a bridge or paths on which D is an ancestor of IV must be blocked by X).

Theorem 5. *If a set of nodes X and a possible instrument IV in internal selection graph G_S^+ satisfy the ignorability criterion relative to D (treatment), and Y (outcome), then $Y_d \perp\!\!\!\perp IV | X, S = 1$.*

This result is proved in Appendix C. We want the instrument to associate with the outcome only along paths that end in a causal path between the treatment and the outcome. This is because the causal paths between the treatment and the outcome are those that we ultimately are interested in studying. So other paths that associate the instrument and the outcome are a problem for instrumental variables approaches. If the instrument were associated with the outcome along some other type of path, we would not be able to disentangle the association between the instrument and the outcome from the association that runs from the instrument to the treatment and then to the outcome along causal paths between the treatment and outcome alone. The latter contains the relationship we want to study, namely the causal relationship between the treatment and the outcome. Conditioning on D or a descendant of D will either block the paths that end in a causal path from the treatment to the outcome or will open non-causal paths between the instrument and the outcome. The ignorability criterion formalizes these ideas. Our ignorability criterion leaves the types of paths between IV and Y that we want to leverage unblocked and requires us to block the types of paths that introduce covariation between IV and Y that can contaminate our analysis.

4.3 Discussion

We now have all the necessary machinery in place to start analyzing how sample selection can threaten or provide opportunities for relevance and ignorability and, hence, instruments. We start in this section by briefly considering some simple examples. Figure 4.1 provides a good starting place. When sample selection is not causally related to any of the other variables in the causal model and we have the canonical instrumental variables graph, as in Figure 4.1(a), we see that we can easily verify relevance and ignorability, where X is the empty set. There is a causal path connecting IV and D, giving relevance. The empty set blocks all paths between IV and Y, except for one that ends in a causal path from $D \rightarrow Y$ and S is not a descendant of D and $D \notin \{X, S\}$, giving ignorability.

Figure 4.1(b,c) both also meet the relevance and ignorability criteria, as the reader can verify for themselves. Figure 4.1(b) is interesting in that the instrumental variables approach here can actually be used to overcome both the unobserved confounding between D and Y from U as well as the sample selection bias between D and Y, even when U, U_1, and U_2 are all unobserved. Figure 4.1(c) is an example of an ancestral instrument that only satisfies the relevance criterion due to the purely statistical association created by sample selection. Perhaps there are opportunities to exploit these types of sample selection mechanisms that have been underappreciated. We will discuss settings like Figure 4.1(b,c) in more detail in a subsequent section.

Figure 4.1(d) is interesting in that it shows that selection on the treatment actually leads to a violation of the ignorability criterion. In simple covariate adjustment approaches (i.e., not instrumental variables), selection based on the treatment is, on its own, not biasing. See Chapter 2 for details. However, in the instrumental variables case, such a selection mechanism clearly does not satisfy the ignorability criterion. This is perhaps the simplest example for which sample selection does not operate in the same way for covariate adjustment approaches and instrumental variables. Researchers should not assume sample selection can be treated similarly in these two approaches. Figure 4.1(e) is a simple example of how, as in simple covariate adjustment approaches, selection on the outcome can violate ignorability. Together, these demonstrate that heuristics from other research designs should not be applied to instrumental variables without thoughtful consideration or the use of formal design-specific criteria, like those in this paper and Chapter 2. We emphasize to the reader that one cannot credibly ascertain the implications of sample selection on an instrumental variables (or any other design) without laying out how sample selection fits into the causal model and using tools like internal selection graphs and our graphical criteria. Less formal approaches will not confer the same assurance that the researcher has not missed some subtle alteration that sample selection makes to the relationships in the data. In the following discussion, we look at more examples aimed at exploring the various ways that sample selection can alter instrumental variables approaches.

4.3.1 Interesting cases

We now consider settings in which the way that sample selection interacts with the instrumental variables design is perhaps under-appreciated or under-discussed. These cases highlight potential opportunities in which to use instrumental variables and also illustrate how sample selection can threaten internal validity of instrumental variables. We hope they are thought-provoking and clarifying.

4.3.1.1 Instrumental variables can be used to recover from sample selection

There are settings in which sample selection can create generalized non-causal paths between the treatment and the outcome, and hence bias designs other than instrumental variables, but for which an instrumental variables design can overcome the sample selection bias. This setting has been recognized elsewhere in the literature (Swanson, 2019). See Figure 4.1(b) for a simple example. The particular form shown in Figure 4.1(b) is commonly called "M-Bias." More generally, an instrumental variables approach could be used to overcome sample selection bias that takes the form of a generalized non-causal paths running between the treatment and outcome created by sample selection (i.e., containing a bridge) that start with an arrow pointing into the treatment. There are really two equivalent ways to look at such settings. One is that instrumental variables is immune to this type of sample selection bias. The other is that instrumental variables is an approach that could be used when this form of sample selection bias is suspected to threaten the validity of simple covariate adjustment approaches. Both views are interesting. The former is useful to know when a researcher plans to employ instrumental variables before considering sample selection. The latter actually presents opportunities for which instrumental variables approaches might be employed. For example, if a researcher suspects that they have a sample selection M-bias problem, they might use an instrumental variables approach to overcome this bias.

4.3.1.2 Ancestral instruments via sample selection

Another interesting case arises when we consider how sample selection can alter relevance. Above, we mentioned the idea of "ancestral" instruments. These are instruments that meet the relevance criterion only when we condition on some covariate(s). (Van Der Zander et al., 2015) It turns out that sample selection can also create ancestral instruments. This is another setting that has been mentioned in the literature (Hernán and Robins, 2020) but is not widely used or discussed and presents additional opportunities for the use of instrumental variables.

A simple version of this appears in Figure 4.1(c). In this setting, U^* is an unobserved but causal instrument. Sample selection creates a purely statistical relationship between IV and U^* in the sample at hand. It is then easy to verify that IV satisfies both the relevance and ignorability criteria. So we can view IV as a proxy for the causal instrument U^*.

While this setting at first sounds quite promising, we urge caution. The nature of this setting makes it likely that there will be violations of ignorability. This setup is only useful if the causal instrument, U^*, is unobserved but is not a common cause to the wrong variables. For instance, ignorability will be violated if U^* is also a parent of Y or U or if there is an unobserved common cause of U^* and Y or various other relationships that might arise in realistic settings. So, while ancestral instruments that arise from sample selection might be intriguing, great care should be taken in evaluating whether ignorability holds for them.

4.3.1.3 Restricting to units that receive two treatments when more exist

Swanson et al. (2015); Ertefaie et al. (2016) discuss the common practice of employing instrumental variables approaches to study how two particular treatments or treatment levels compare, where the sample is limited to units receiving these two treatments, but where more than two treatments are possible. Such studies suffer from selection on the treatment, which, as we can see in Figure 4.1(d), can violate ignorability and bias effect estimates. However, the "ensuing selection bias that occurs due to this restriction has gone relatively unnoticed" and is "pervasive." (Swanson et al., 2015) We echo that this and similar practices are a problem as they violate ignorability.

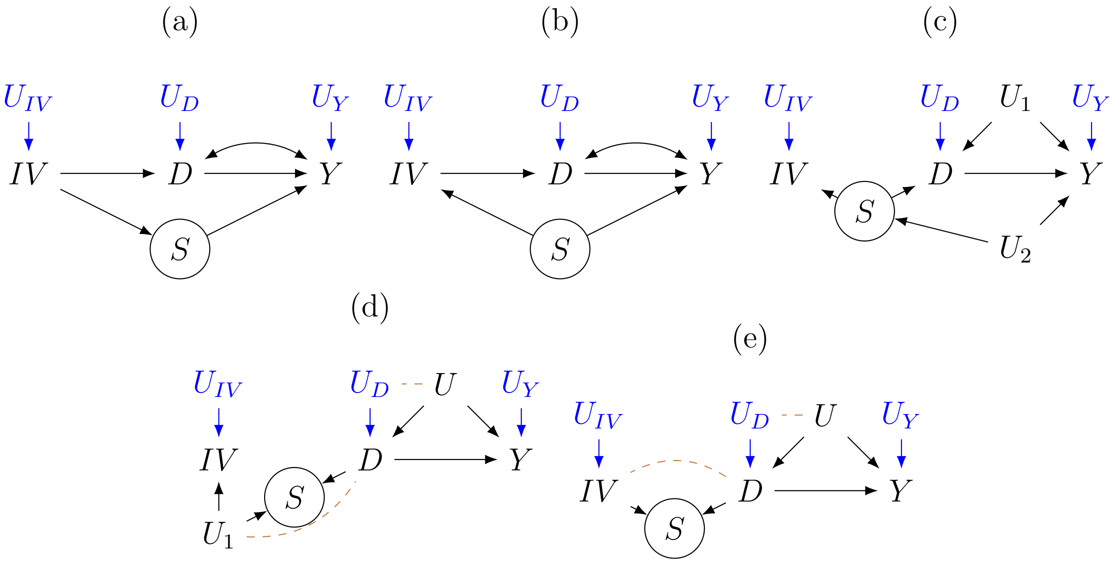
4.3.1.4 Randomized experiments can have sample selection problems

Random experiments are often held as the gold-standard for causal inference. However, they are not impervious to threats to validity. Often experiments, especially when related to human subjects, have non-compliance with assigned treatments. This means that participants choose not to, say, take a drug when they've been prescribed it. Which participants choose not to comply could have common causes with the outcome, confounding the treatment-outcome relationship. Instrumental variables can be used to address this non-compliance by using assigned treatment as an instrument for whether or not someone is actually treated (e.g., actually takes the drug). A second threat to the validity of randomized experiments is differential attrition from the study. This means that some participants or units drop out of the study and so complete data is not available for all participants. When both non-compliance and attrition both occur, a researcher might be in a scenario where, despite randomly assigning treatment, they are attempting to use instrumental variables while also worrying about sample selection. The simplest forms of this would be captured by Figure 4.1(d) and (e). Despite the randomization of treatment assignment, there are violations of ignorability and effect estimates will be biased. Alternatively, if attrition is based on the instrument and there is a common cause of attrition and the outcome, we might also have a violation of ignorability. See Montgomery et al. (2018) for a very useful discussion about post-treatment conditioning and selection in randomized experiments in political science and the bias that can result.

4.3.1.5 Sample selection can make and break both ignorability and relevance

As we saw with ancestral instruments, sample selection can sometimes help provide relevance. This is also true for ignorability. See Figure 4.2(a) and (b). However, it can also create ignorability while breaking relevance, as in Figure 4.2(c). Similarly, sample selection can also create relevance while breaking ignorability. See Figure 4.2(d) and (e). At first, glance these last two example may seem like they satisfy ignorability. While paths like $IV \cdots D \to Y$ are allowed since they contain $D \to Y$, $IV \cdots D$ also creates the path $IV \cdots D \leftarrow U \to Y$, which breaks ignorability.

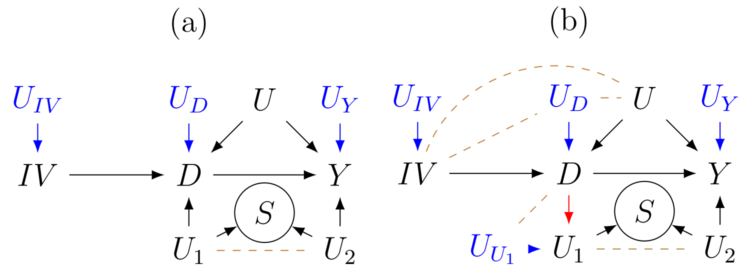
Figure 4.2: Sample selection can make and break both ignorability and relevance



4.3.1.6 Small changes to the causal graph matter

We've stressed the importance of including a sample selection node in every causal graph. We now stress that great care must be taken in constructing causal graphs containing sample selection node. In Figure 4.3 we see the same graph as in Figure 4.1(b) but where we flip the direction of just one edge. The conclusions for the two variations on the graph are opposites. In Figure 4.3(a) we satisfy the ignorability criterion but in Figure 4.3(b) we do not. Researchers need to be careful about the details of the causal graph that they are studying. Including a sample selection node in every causal model and careful consideration of the sample selection mechanisms is required to determine the threat that sample selection poses to internal validity and what, if anything, might be able to be done. There is also potential for users to intentionally or unintentionally favor one graph over another very similar graph in order to show that ignorability holds. These are difficult but inherent problems in causal study and good-faith efforts to do credible causal inference should spend ample time defending the specific causal model being analyzed.

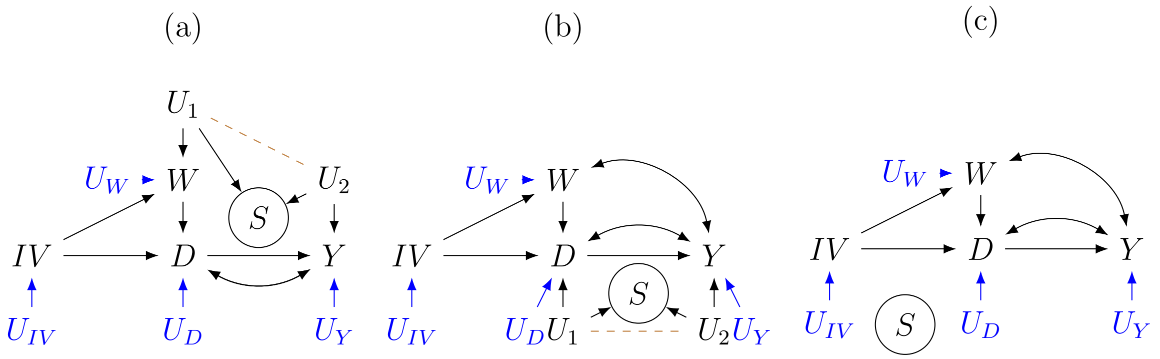
Figure 4.3: Small changes to the causal graph matter



4.3.1.7 Blocking non-causal paths between treatment and outcome can invalidate instruments

The final interesting case we consider again cautions against applying heuristics from simple covariate adjustment approaches (or other designs) in the instrumental variables setting. In both simple covariate adjustment and instrumental variables, we want to learn about the causal effect of the treatment on the outcome. In a covariate adjustment setting, we attempt to de-confound the true treatment effect by blocking non-causal paths between the treatment and the outcome. We can reduce the bias of our estimates by limiting the number of non-causal paths that run between the treatment and the outcome. Doing so helps to make the treated group comparable to the untreated group (assuming a binary treatment), making a comparison of these groups useful for causal analysis. We employ instrumental variables approaches only when we cannot block *all* the non-causal paths between the treatment and the outcome. And when this is the case, there are settings in which blocking non-causal paths between the treatment and the outcome can actually lead to violations of the ignorability we need for instruments. This can happen even if we would have had a valid instrument before blocking the non-causal paths between the treatment and the outcome. See Figure 4.4 for three examples. Note that not all of these examples hinge on sample selection creating bias. In these examples, conditioning on W would block a non-causal path between the treatment and the outcome but open others between IV and Y, even though no open non-causal paths existed between IV and Y before conditioning on W.

Figure 4.4: Blocking non-causal paths between treatment and outcome can invalidate instruments



We hope the settings discussed in this section provide some further insight into how sample selection can effect instrumental variables. We again stress that, without incorporation of the sample selection mechanism into the causal model and use of a formal framework like that presented here, there is no reliable way to determine how sample selection might alter relevance or ignorability for your specific instrumental variables application.

4.3.2 Lessons

After considering the above examples, let us review key lessons.

- Only the incorporation of the sample selection mechanism into the causal model can reliably lead to correct conclusions about how sample selection might effect instrumental variables. Moreover, graphical analysis and a formal framework for the analysis of sample selection can greatly reduce the burden on researcher in analysis of how sample selection alters their instrumental variables approach.
- Informal applications of simple heuristics related to sample selection can be misleading and should be avoided. Further, we should not use heuristics from other research designs, like simple covariate adjustment, in the instrumental variables setting. These might not apply (e.g., like selection on the treatment being non-biasing) and can result

in unreliable conclusions.

- Sample selection can influence instrumental variables, even when it is not a collider.
- When sample selection manifests as attrition or some other post-treatment type of selection, randomization of treatment or instrument assignment does not automatically ameliorate problems of sample selection even for internal validity.
- Sample selection does not always present a problem.
- Selection on the outcome is usually a problem for instrumental variables.
- Post-treatment selection is typically a problem on its own.
- Indirect association between selection and the outcome or selection and the treatment are typically not problems on their own and also typically not when they appear together for instrumental variables.
- There are various ways that sample selection can threaten relevance and ignorability.
- There are also opportunities presented by sample selection for instrumental varibles as well as by instrumental variables for sample selection.

In conclusion, instrumental variables approaches leverage specific types of variation between the instrument and treatment and the instrument and outcome to identify causal effects of the treatment on the outcome. We've seen how these associations can be altered in a non-randomly selected sample in a variety of ways. Sample selection can create many wrinkles in an instrumental variables analysis but is not always problematic. However, the already high bar of finding a good instrument is only made higher by responsibly considering how sample selection can influence instruments. We hope that our discussion provides clarity and credibility to researchers that hope to use instrumental variables.

CHAPTER 5

Partial identification leveraging imperfect placebos

In this chapter, we shift our focus from sample selection to unobserved confounding. As in previous chapters, however, the tools we develop can be applied to address either or both threats to internal validity. Unobserved confounding plagues many attempts to draw causal inferences from observational data. When we suspect unobserved confounding exists, we often look for additional sources of information that can aid us in identifying causal effects of interest. This may take the form of an instrumental variable or some sort of discontinuity in the assignment of which units get treated. Researchers can also seek out observed variables that are "similar to" the treatment or outcome in the form of placebo treatments or placebo outcomes (also known as "negative control outcomes" and "negative exposure controls").

We focus here on how to make use of such placebo treatment and outcomes variables to improve causal inferences. Existing approaches using such information typically focus on detecting bias or on point identification (i.e., bias correction), assuming (i) that these are "perfect placebos" (treatment has precisely zero effect on a placebo outcome, or placebo treatments have precisely zero effects on the outcome of interest) and (ii) that the placebo outcome/treatment suffers the same amount of confounding as does the real treatment-outcome relationship on some scale ("equiconfounding"). As discussed below, difference-in-difference can be understood as one such approach. Another recent innovation of this kind is proximal causal inference (Miao et al., 2020; Tchetgen et al., 2020). In this approach, point identification is possible, at the expense of (i) assuming that placebos are "perfect," (ii) requiring a placebo treatment and a placebo outcome, and (iii) requiring equiconfounding on a transformation of a placebo outcome.[1]

We start from the position that assumptions such as exact equiconfounding and/or having "perfect" placebos are often indefensible in practice. Nevertheless, information from outside the data may allow us to reasonably bound a set of parameters pertaining to these assumptions. Specifically, consider a set of assumptions about (i) the relative confounding felt by the treatment-outcome relationship compared to that in the placebo treatment to real outcome or real treatment to placebo outcome relationship; and (ii) the degree of "imperfection" in each placebo. For any range of assumptions on these that cannot be ruled out by argumentation, there is a set of effect estimates that consequently must be entertained as possible. In short, we leverage the assumptions investigators may be willing to make about imperfect placebo treatment and outcomes and about the relative confoundedness of placebos versus "real" treatments and outcomes, to produce partial identification of parameters in a linear model.

In what follows, we begin by discussing causal structures in general that involve placebo treatments and outcomes, including cases where these labels become ambiguous. We then develop a simple framework for working with a single placebo outcome, relaxing assumptions that treatment has no effect on this placebo outcome and postulating different strengths of confounding between treatment and the real outcome vs. treatment and the placebo outcome. We can then extend similar tools and mechanics to work with placebo treatments, and finally cases with both placebo treatments and outcomes.[2] We then explore at some length how conventional difference-in-differences (DiD) can be seen as a special case, where

[1] Tchetgen et al. (2023a) presents a related approach that requires a single "perfect" placebo outcome. That is, the placebo, W, must satisfy $W_d = W$, $W \perp\!\!\!\perp Y_{D=0}$, and $W \perp\!\!\!\perp D|Y_{D=0}$, where D is the treatment, Y is the outcome, and $Y_{D=0}$ is the value for the Y if D were to be, perhaps counterfactually, to take the value $D = 0$. The authors state that "the key assumption that conditioning on the treatment-free potential outcome, would in principle make the NCO or outcome proxy irrelevant to treatment mechanism (sic) is ultimately untestable, and may in certain settings not hold exactly." Our approach can be seen as relaxing the these conditions. Tchetgen et al. (2023a) also require more technical conditions related to solutions to integral equations. For example, they require that there is a solution $b(W)$ to the equation $Y_{D=0} = \mathbb{E}[b(W)|Y_{D=0}]$, when the previous conditions are also met. This condition can be thought of as a generalization of measurement error.

[2] We also explore generalizations to partially linear and non-parametric settings in Appendix D.2.

the pre-treatment outcome is a (perfect) placebo outcome, and the parallel trends assumption is a strict equiconfounding assumption on a specific scale. The tools we develop, by relaxing the equiconfounding assumption, profitably relax the parallel trends assumption and offer a partial identification approach to produce a range of estimates the investigator is prepared to defend. Finally, we provide two applications, employing an `R` package we developed. First, we look to observational data designed to examine the impact of the National Supported Work Demonstration, a job training program in the 1970s. A pre-treatment measure of pariticpants' income serves as a placebo outcome. Assumptions are then required on how strongly pre-treatment income is confounded compared to how strongly treatment and subsequent income are confounded. Under arguably defensible assumptions, the resulting range of estimates is informative and includes the benchmark value obtained from a randomized trial with the same treatment group. Second, we show that for a likely range of relative levels of confounding, the Zika virus decreased birth rates in Brazil in 2016, using 2014 birth rates as a placebo.

5.1 Leveraging placebos

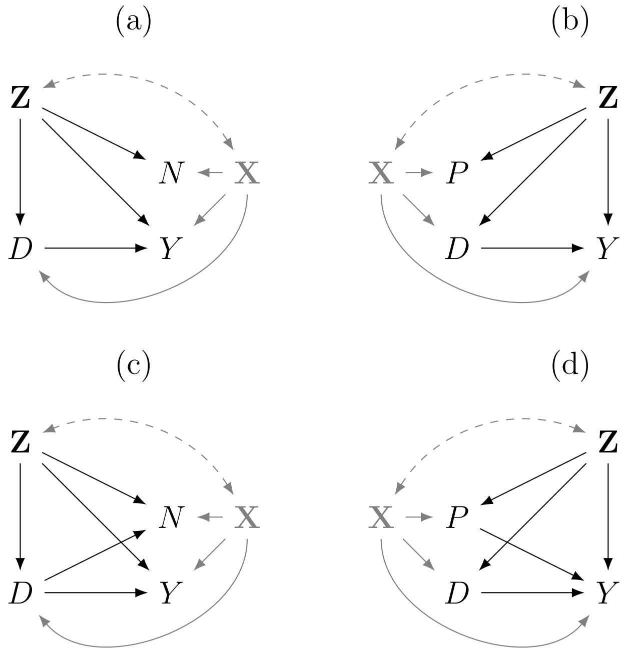
5.1.1 Placebo causal structures

We begin by ensuring clarity regarding what we mean by placebo treatments and outcomes, perfect and imperfect, by reference to their position in a causal model as represented by a directed acyclic graphs (DAGs) (Pearl, 1995). To start with "perfect" placebos, two simple examples are shown in Figure 5.1(a,b). In this figure N is a placebo outcome and P is a placebo treatment; however, Figure 5.1(b) is essentially a mirror image of Figure 5.1(a) with N relabelled as P. What we consider a placebo treatment versus what what we consider a placebo outcome will often have this "Necker cube" quality (Necker, 1832).[3] Depending on the substantive context, the same causal graph might arise in two settings, one in which

[3] A Necker cube is a simple optical illusion in which the perceived orientation of a a two-dimensional drawing of a cube can continuously switch back and forth between two distinct possibilities.

the placebo is viewed as a placebo treatment and one in which the placebo is viewed as a placebo outcome. We explore this more in Section 5.2 and discuss how the perspective taken by researchers should influence the methods they use. To call a placebo "imperfect" is to allow for an additional causal effect, and so always implies an additional edge in the causal structure. Figure 5.1(c,d) contains simple examples of such *imperfect* placebos. Figure 5.1(c) adds the $D \to N$ edge to Figure 5.1(a), while 5.1(d) adds the $P \to Y$ edge to Figure 5.1(b).

Figure 5.1: DAGs with a placebo outcome, N, and with a placebo treatment, P. D is treatment, Y is outcome, $\mathbf{Z}$ contains unobserved confounders, and $\mathbf{X}$ contains observed covariates.



5.1.2 Placebo outcomes

Our proposed framework for using information from placebo treatments and outcomes relies the omitted variable bias analysis for linear regression. However, we also explore relaxations to this parametric approach in Appendix D.2, in which we illustrate similar methods adapted to partially linear and non-parametric estimation of treatment effects. In either case, we show

how the omitted variable bias paradigm can be augmented to incorporate information from placebo treatments and outcomes to make causal progress with unequal levels of confounding for a placebo treatment/outcome compared to the true treatment-outcome relationship and with imperfect placebos. We emphasize this approach as a partial identification strategy, but it can equivalently be seen as providing sensitivity analysis for more traditional placebo approaches. We start by discussing placebo outcomes, then discuss placebo treatments, and finally discuss "double placebos."

We start our discussion of placebo outcomes by leveraging the traditional omitted variable bias framework. (Angrist and Pischke, 2008; Cinelli and Hazlett, 2020; Hansen, 2022) Suppose that we want to estimate the "long" regression of Y on $D, \mathbf{X}$, and $\mathbf{Z}$ in Equation 5.1. In particular, we are interested in the coefficient $\beta_{Y\sim D|Z,X}$ as a measure of the causal effect of D on Y or as an approximation to the causal effect of D on Y. $\mathbf{X}$ and $\mathbf{Z}$ are each vectors of covariates. We can consider $Z_{(\text{Y.DX})} = \mathbf{Z}\beta_{Y\sim Z|D,X}$ to be the linear combination of $\mathbf{Z}$ that "de-confounds" the $D \rightarrow Y$ relationship in the same way that including $\mathbf{Z}$ in Equation 5.1 does.[4] That is $\beta_{Y\sim D|Z,X} = \beta_{Y\sim D|Z_{(\text{Y.DX})},X}$ from Equation 5.1 and 5.2.[5]

$$Y = \beta_{Y\sim D|Z,X} D + \mathbf{X}\beta_{Y\sim X|D,Z} + \mathbf{Z}\beta_{Y\sim Z|D,X} + \epsilon_l \tag{5.1}$$

$$Y = \beta_{Y\sim D|Z_{(\text{Y.DX})},X} D + \mathbf{X}\beta_{Y\sim X|D,Z_{(\text{Y.DX})}} + \beta_{Y\sim Z_{(\text{Y.DX})}|D,X} Z_{(\text{Y.DX})} + \epsilon_l \tag{5.2}$$

In practice, we typically do not observe all the covariates that would allow us to estimate or approximate the causal effect of D on Y by estimating this long regression. Therefore, let us assume that $\mathbf{Z}$ is unobserved. We will refer to these as "unobserved confounders." Since $\mathbf{Z}$

[4] Note that in this section and others, we will refer to including $\mathbf{Z}$ in the long regression as "de-confounding". Readers should recognize that, in reality, we are only partialling out the linear relationships with $\mathbf{Z}$. This is, therefore, a linear approximation to full non-parametric elimination of confounding from $\mathbf{Z}$. On the other hand, $\mathbf{Z}$ can be assumed to contain arbitrary functions of the relevant variables, making this very flexible for de-confounding.

[5] Indeed $R^2_{Y\sim Z_{(\text{Y.DX})}|D,X}$ will also equal $R^2_{Y\sim Z|D,X}$. But $R^2_{D\sim Z_{(\text{Y.DX})}|X} \leq R^2_{D\sim Z|X}$, since $Z_{(\text{Y.DX})}$ is chosen to maximize the R^2 between $\mathbf{Z}$ and Y not $\mathbf{Z}$ and D. (Cinelli and Hazlett, 2020)

are unobserved, we must estimate the "short" regression of Y on D and $\mathbf{X}$ in Equation 5.3, rather than the desired long regression. $\mathbf{Z}$ is "omitted" from the regression, though we wish we could have included it.

$$Y = \beta_{Y\sim D|X} D + \mathbf{X}\beta_{Y\sim X|D} + \epsilon_s \tag{5.3}$$

Traditional omitted variable bias analysis tells us that $\beta_{Y\sim D|X} - \beta_{Y\sim D|Z_{(\text{Y.DX})},X}$ can be characterised as the product of two regression coefficients, as in Equation 5.4. $\beta_{Y\sim Z_{(\text{Y.DX})}|D,X}$ captures the relationship between Y and $Z_{(\text{Y.DX})}$ after linearly partialling out D and X. $\beta_{Z_{(\text{Y.DX})}\sim D|X}$ captures the relationship between D and $Z_{(\text{Y.DX})}$ after linearly partialling out X.

$$\text{bias}_{(\text{YD.X})} \triangleq \beta_{Y\sim D|X} - \beta_{Y\sim D|Z_{(\text{Y.DX})},X} = \beta_{Y\sim Z_{(\text{Y.DX})}|D,X}\beta_{Z_{(\text{Y.DX})}\sim D|X} \tag{5.4}$$

These components alone are enough to build a powerful partial identification framework, as discussed in Cinelli and Hazlett (2020). But can more be done when there also exists an observed placebo outcome, N? If we think that the relationship between D and N suffers from a similar level of unobserved confounding as does the relationship between D and Y, then how might we leverage N to understand $\text{bias}_{(\text{YD.X})}$ and, hence, $\beta_{Y\sim D|Z,X} = \beta_{Y\sim D|Z_{(\text{Y.DX})},X}$? Consider a long and short regression with N as the outcome and with D as the treatment. The long regression of N on $D, \mathbf{X}$, and $\mathbf{Z}$ can be found in Equation 5.5. We can again consider $Z_{(\text{N.DX})} = \mathbf{Z}\beta_{N\sim Z|D,X}$ to be the linear combination of $\mathbf{Z}$ that "de-confounds" the $D \rightarrow N$ relationship in the same way that including $\mathbf{Z}$ in Equation 5.5 does. Note that we can consider $\mathbf{Z}$ to be a rich enough set of variables to de-confound both the $D \rightarrow Y$ and the $D \rightarrow N$ relationships. Thus, $\beta_{N\sim D|Z,X} = \beta_{N\sim D|Z_{(\text{N.DX})},X}$ from Equation 5.5 and 5.6. It is important to note that $Z_{(\text{N.DX})}$ will not equal $Z_{(\text{Y.DX})}$ when $\beta_{N\sim Z|D,X} \neq \beta_{Y\sim Z|D,X}$. The short regression of N on just D and $\mathbf{X}$ can be found in Equation 5.7. As with Y, we are only able to consider estimating the short regression for N, as $\mathbf{Z}$ is unobserved. Here too $\beta_{N\sim D|X} - \beta_{N\sim D|Z_{(\text{N.DX})},X}$ can be characterised as in Equation 5.8.

$$N = \beta_{N\sim D|Z,X} D + \mathbf{X}\beta_{N\sim X|D,Z} + \mathbf{Z}\beta_{N\sim Z|D,X} + \xi_l \tag{5.5}$$

$$N = \beta_{N\sim D|Z_{(\text{N.DX})},X} D + \mathbf{X}\beta_{Y\sim X|D,Z_{(\text{N.DX})}} + \beta_{N\sim Z_{(\text{N.DX})}|D,X} Z_{(\text{N.DX})} + \xi_l \tag{5.6}$$

$$N = \beta_{N\sim D|X} D + \mathbf{X}\beta_{N\sim X|D} + \xi_s \tag{5.7}$$

$$\text{bias}_{(\text{ND.X})} \triangleq \beta_{N\sim D|X} - \beta_{N\sim D|Z_{(\text{N.DX})},X} = \beta_{N\sim Z_{(\text{N.DX})}|D,X}\beta_{Z_{(\text{N.DX})}\sim D|X} \tag{5.8}$$

From Equations 5.4 and 5.8, it is easy to see how to leverage information about the placebo outcome, N. There is some $m \in \mathbb{R}$ such that $\text{bias}_{(\text{YD.X})} = m \times \text{bias}_{(\text{ND.X})}$. Thus, we can write Equations 5.9 and 5.10; in these expressions, we substitute $\beta_{N\sim D|Z,X}$ for $\beta_{N\sim D|Z_{(\text{N.DX})},X}$ and $\beta_{Y\sim D|Z,X}$ for $\beta_{Y\sim D|Z_{(\text{Y.DX})},X}$, since they are equal. This can be thought of as using the placebo outcome to re-express the confounding path $D \leftarrow Z \rightarrow Y$ in Figure 5.1(a).

$$\text{bias}_{(\text{YD.X})} = \beta_{Y\sim D|X} - \beta_{Y\sim D|Z,X} = m \times \left(\beta_{N\sim D|X} - \beta_{N\sim D|Z,X}\right) \tag{5.9}$$

$$\iff \beta_{Y\sim D|Z,X} = \beta_{Y\sim D|X} - m \times \left(\beta_{N\sim D|X} - \beta_{N\sim D|Z,X}\right) \tag{5.10}$$

Equation 5.9 is an expression for $\text{bias}_{(\text{YD.X})}$ in terms of $m \triangleq \frac{\beta_{Y\sim Z_{(\text{Y.DX})}|D,X}\beta_{Z_{(\text{Y.DX})}\sim D|X}}{\beta_{N\sim Z_{(\text{N.DX})}|D,X}\beta_{Z_{(\text{N.DX})}\sim D|X}} = \frac{\text{bias}_{(\text{YD.X})}}{\text{bias}_{(\text{ND.X})}}$, $\beta_{N\sim D|Z_{(\text{N.DX})},X}$, and $\beta_{N\sim D|X}$. We have a similar expression for $\beta_{Y\sim D|Z_{(\text{Y.DX})},X}$ in Equation 5.10, our target parameter, that also includes $\beta_{Y\sim D|X}$. We can estimate $\beta_{N\sim D|X}$ and $\beta_{Y\sim D|X}$ from the data.[6] The remaining terms in these expressions can be treated as parameters that can be specified by investigators. Thus, we can use Equation 5.10 for partial identification of $\beta_{Y\sim D|Z,X}$ by establishing plausible ranges of values for m and $\beta_{N\sim D|Z,X}$. m captures the level of relative confounding. This can be thought of as comparing the strength of the $D \leftarrow Z \rightarrow Y$ relationship to the $D \leftarrow Z \rightarrow N$ relationship in Figure 5.1(a).

[6] Note that we do not need the placebo outcome to be observed in the same dataset or for the same units as the primary outcome to estimate $\beta_{N\sim D|X}$ and $\beta_{Y\sim D|X}$.

$\beta_{N\sim D|Z,X} = \beta_{N\sim D|Z_{(\text{N.DX})},X}$ captures any direct causal relationship between D and N, like the $D \rightarrow N$ path in Figure 5.1(c). Therefore, this $\beta_{N\sim D|Z,X}$ parameter allows for "imperfect" placebo outcomes that are directly affected by the treatment. Since both N and D are observed variables, their relationship conditional on Z, X should be able to be specified as a regression coefficient. In many cases differing scales for N and Y might lead to difficulty in interpreting or reasoning about the values for m. Can we do better?

Re-expressing m Let us consider a reparameterization of m in terms of scale-independent partial-correlation parameters, rather than regression coefficients. This will allow us to circumvent the scale issues we just mentioned. To do this, we use the reparameterizations of omitted variable bias similar to that discussed in Cinelli and Hazlett (2020) shown in Equation 5.11.

$$\begin{aligned}
\text{bias}_{(\text{YD.X})} &= \beta_{Y\sim D|X} - \beta_{Y\sim D|Z_{(\text{Y.DX})},X} = R_{Y\sim Z_{(\text{Y.DX})}|D,X}\, f_{D\sim Z_{(\text{Y.DX})}|X}(\text{SD}(Y^{\perp D,X})/\text{SD}(D^{\perp X}))\\
\text{bias}_{(\text{ND.X})} &= \beta_{N\sim D|X} - \beta_{N\sim D|Z_{(\text{N.DX})},X} = R_{N\sim Z_{(\text{N.DX})}|D,X}\, f_{D\sim Z_{(\text{N.DX})}|X}(\text{SD}(N^{\perp D,X})/\text{SD}(D^{\perp X}))
\end{aligned} \tag{5.11}$$

$$m = \frac{\text{bias}_{(\text{YD.X})}}{\text{bias}_{(\text{ND.X})}} = \frac{R_{Y\sim Z_{(\text{Y.DX})}|D,X}\, f_{D\sim Z_{(\text{Y.DX})}|X}}{R_{N\sim Z_{(\text{N.DX})}|D,X}\, f_{D\sim Z_{(\text{N.DX})}|X}} \times \frac{\text{SD}(Y^{\perp D,X})}{\text{SD}(N^{\perp D,X})} = k \times \frac{\text{SD}(Y^{\perp D,X})}{\text{SD}(N^{\perp D,X})} \tag{5.12}$$

$$\begin{aligned}
k \triangleq \frac{R_{Y\sim Z_{(\text{Y.DX})}|D,X}\, f_{D\sim Z_{(\text{Y.DX})}|X}}{R_{N\sim Z_{(\text{N.DX})}|D,X}\, f_{D\sim Z_{(\text{N.DX})}|X}} &= \frac{\text{bias}_{(\text{YD.X})}}{\text{bias}_{(\text{ND.X})}} \times \frac{1}{\text{SF}}\\
&= \frac{\text{bias}_{(\text{YD.X})} \times (1/\text{SF})}{\text{bias}_{(\text{ND.X})}} = \frac{\text{bias}_{(\text{YD.X})}}{\text{bias}_{(\text{ND.X})} \times \text{SF}}\\
\text{where } \text{SF} &= \frac{\text{SD}(Y^{\perp D,X})}{\text{SD}(N^{\perp D,X})}
\end{aligned} \tag{5.13}$$

Equation 5.12 provides an expression for m in terms of k[7] and SF. SF ("scale factor")

[7] If partial R^2 are more intuitive, users can substitute in $k = \text{sign}(k) \times \sqrt{k^2}$ and consider $\text{sign}(k) =$

captures the scale issues and deals only with observables. The numerator and denominator are moments of the residuals from the two short regressions. k is a scaled version of the ratio of $\text{bias}_{(\text{YD.X})}$ over $\text{bias}_{(\text{ND.X})}$. This is a scale independent measure of the level of confounding of the Y, D relationship relative to the level of confounding of the N, D relationship. Alternatively put, it is the level of unobserved confounding in the Y, D relationship as a percentage of the level of unobserved confounding in the N, D relationship, after resclaing to account for possible differences in the scale of Y and N. If $\text{SD}(N^{\perp D,X}) = \text{SD}(Y^{\perp D,X})$, then k is simply the ratio of biases. Quantities like $R_{Y\sim Z_{(\text{Y.DX})}|D,X} f_{D\sim Z_{(\text{Y.DX})}|X}$ and $R_{N\sim Z_{(\text{N.DX})}|D,X} f_{D\sim Z_{(\text{N.DX})}|X}$ are called "bias factors" in Cinelli and Hazlett (2020). $f_{D\sim Z|X}$ is Cohen's f which equals $\frac{R}{\sqrt{1-R^2}}$. We can also re-write the bias-factors as $R_{Y\sim Z|D,X} f_{D\sim Z|X} = f_{Y\sim D|X} - f_{Y\sim D|Z,X}\sqrt{\frac{1-R^2_{Y\sim Z|D,X}}{1-R^2_{D\sim Z|X}}}$. See Appendix D.3 for the derivation. Equations 5.9 and 5.10 can be rewritten as Equation 5.14 and 5.15.

$$\text{bias}_{(\text{YD.X})} = k \times \left(\beta_{N\sim D|X} - \beta_{N\sim D|Z,X}\right) \times \frac{\text{SD}(Y^{\perp D,X})}{\text{SD}(N^{\perp D,X})} \tag{5.14}$$

$$\beta_{Y\sim D|Z,X} = \beta_{Y\sim D|X} - k \times \left(\beta_{N\sim D|X} - \beta_{N\sim D|Z,X}\right) \times \frac{\text{SD}(Y^{\perp D,X})}{\text{SD}(N^{\perp D,X})} \tag{5.15}$$

Estimator and partial identification Equation 5.15 provides us with a moment estimator for $\beta_{Y\sim D|Z,X}$ in Equation 5.16.[8][9]

$\text{sign}\left(\frac{R_{Y\sim Z_{(\text{Y.DX})}|D,X} f_{D\sim Z_{(\text{Y.DX})}|X}}{R_{N\sim Z_{(\text{N.DX})}|D,X} f_{D\sim Z_{(\text{N.DX})}|X}}\right)$ and $k^2 = \frac{R^2_{Y\sim Z_{(\text{Y.DX})}|D,X} f^2_{D\sim Z_{(\text{Y.DX})}|X}}{R^2_{N\sim Z_{(\text{N.DX})}|D,X} f^2_{D\sim Z_{(\text{N.DX})}|X}}$. The denominator of k cannot equal zero. This will be the case for any useful placebo outcome. Otherwise, the placebo outcome will not have a confounded relationship with the treatment and not have any useful information that we can leverage to understand the confounding of the relationship between the treatment and the actual outcome.

[8] We follow a similar presentation of our estimators and approaches to inference as in Zhang and Ding (2022).

[9] It is also possible to show that $\frac{\widehat{\text{SD}}(Y^{\perp D,X})}{\widehat{\text{SD}}(N^{\perp D,X})} = \frac{\left[\frac{\widehat{\text{SD}}(Y^{\perp D,X})}{\widehat{\text{SD}}(D^{\perp X})}\right]}{\left[\frac{\widehat{\text{SD}}(N^{\perp D,X})}{\widehat{\text{SD}}(D^{\perp X})}\right]} = \frac{\text{se}(\hat{\beta}_{Y\sim D|X})}{\text{se}(\hat{\beta}_{N\sim D|X})} \times \sqrt{\frac{\text{df}_Y}{\text{df}_N}}$, which is an expression containing only commonly reported regression summary statistics from the two short regressions that are estimable from the observed data. $\text{se}(\hat{\beta}_{Y\sim D|X})$ and $\text{se}(\hat{\beta}_{N\sim D|X})$ are the estimated standard errors from standard regression estimation of the short regressions. df_Y and df_N are the degrees of freedom from the estimated short regressions.

$$\hat{\beta}_{Y\sim D|Z,X} = \hat{\beta}_{Y\sim D|X} - k \times \left(\hat{\beta}_{N\sim D|X} - \beta_{N\sim D|Z,X}\right) \times \frac{\|Y^{\perp D,X}\|_2}{\|N^{\perp D,X}\|_2} \tag{5.16}$$

Given k and $\beta_{N\sim D|Z,X}$, the estimator $\hat{\beta}_{Y\sim D|Z,X}$ from Equation 5.16 is consistent for $\beta_{Y\sim D|Z,X}$ and asymptotically normal. We will discuss estimating standard errors in Section 5.1.5. Using Equation 5.16 in a partial identification framework, we can reason about plausible ranges of values for $\beta_{N\sim D|Z,X}$ and k to arrive at a range for plausible estimates of $\beta_{Y\sim D|Z,X}$. In this way, we can partially identify $\beta_{Y\sim D|Z,X}$. In the special cases when there is only a single omitted variable, $Z_{(\text{N.DX})} = Z_{(\text{Y.DX})}$, or $f_{D\sim Z_{(\text{N.DX})}|X} = f_{D\sim Z_{(\text{Y.DX})}|X}$, then k can also be written as $k = \frac{R_{Y\sim Z|D,X}}{R_{N\sim Z|D,X}}$. This can be thought of as comparing the strength of the $Z \rightarrow Y$ "arm" of the confounding path $D \leftarrow Z \rightarrow Y$ to the relative strength of the $Z \rightarrow N$ "arm" of the confounding path $D \leftarrow Z \rightarrow N$ in Figure 5.1(a).

5.1.3 Placebo treatments

It is possible to use similar machinery to leverage imperfect placebo treatments rather than imperfect placebo outcomes in a partial identification framework for effects of actual treatments on outcomes. Suppose that we want to run the long regression for Y on D (Equation 5.1) but, again, $\mathbf{Z}$ is an unobserved vector of variables that we would have liked to include in the regression but cannot. Suppose also that we have a placebo treatment, P, for which we believe the confounding of P and Y from $\mathbf{Z}$ is similar to the confounding of D and Y from $\mathbf{Z}$. We might consider running two separate regressions as we did for placebo outcomes. However, that approach is not easily adapted to imperfect placebo treatments (i.e., those for which there is a direct relationship between P and Y).[10] Instead, suppose that P is not a descendant of D, D is not a descendant of P, P is a neutral control for the effect of D on Y,

[10] Suppose we have a causal model like Figure 5.1(d). $\beta_{Y\sim Z|P,X}$ (or $R_{Y\sim Z|P,X}$) will capture the causal relationship $D \rightarrow Y$, since the path $Z \rightarrow D \rightarrow Y$ will remain open conditional on P and X alone. Expressing the bias in the regression coefficient on D from the short regression using a two regression approach like we used for placebo outcomes would, therefore, require us to reason about the causal effect of interest itself, which is not useful.

and P is a neutral control for D on Y, conditional on Z and X.[11] Further, suppose that we want to run the "long" regression of Y on $D, P, \mathbf{X}$, and $\mathbf{Z}$ in Equation 5.17. In particular, we are interested in the coefficient $\beta_{Y\sim D|P,Z,X}$ as a measure of the causal effect of D on Y or as an approximation to the causal effect of D on Y. However, since $\mathbf{Z}$ is unobserved, we must run the "short" regression of Y on D, P and $\mathbf{X}$ in Equation 5.18, rather than the desired long regression.

$$Y = \beta_{Y\sim D|P,Z,X} D + \beta_{Y\sim P|D,Z,X} P + \mathbf{X}\beta_{Y\sim X|D,P,Z} + \mathbf{Z}\beta_{Y\sim Z|D,P,X} + \epsilon_l \tag{5.17}$$

$$Y = \beta_{Y\sim D|P,X} D + \beta_{Y\sim P|D,X} P + \mathbf{X}\beta_{Y\sim X|D,P} + \epsilon_s \tag{5.18}$$

Following a similar omitted variables based approach as we took for placebo outcomes (see Equation 5.19), we can arrive at similar partial identification expressions for the placebo treatment case.

[11] We require that D and P are neutral controls for each other in that they do not change the regression coefficients for the outcome Y when we include them in the same regression. This will simplify interpretation. If a user is concerned that there could be a path opened or closed by conditioning on D or P, we suggest drawing the DAG to check. See Section 5.2 for further discussion.

$$\begin{aligned}
&\text{bias}_{(\text{YD.PX})} = \beta_{Y\sim D|P,X} - \beta_{Y\sim D|P,Z_{(\text{Y.DPX})},X} \\
&= R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{D\sim Z_{(\text{Y.DPX})}|P,X} (\text{SD}(Y^{\perp D,P,X})/\text{SD}(D^{\perp P,X})) \\
&\text{bias}_{(\text{YP.DX})} = \beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\text{Y.DPX})},X} \\
&= R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{P\sim Z_{(\text{Y.DPX})}|D,X} (\text{SD}(Y^{\perp D,P,X})/\text{SD}(P^{\perp D,X})) \\
&\text{bias}_{(\text{YD.PX})} = m \times \text{bias}_{(\text{YP.DX})} \\
&\implies \beta_{Y\sim D|P,Z_{(\text{Y.DPX})},X} = \beta_{Y\sim D|P,X} - m \times (\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\text{Y.DPX})},X}) \\
&m = \frac{\text{bias}_{(\text{YD.PX})}}{\text{bias}_{(\text{YP.DX})}} = \frac{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{D\sim Z_{(\text{Y.DPX})}|P,X}}{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{P\sim Z_{(\text{Y.DPX})}|D,X}} \times \frac{\text{SD}(P^{\perp D,X})}{\text{SD}(D^{\perp P,X})} = k \times \frac{\text{SD}(P^{\perp D,X})}{\text{SD}(D^{\perp P,X})}
\end{aligned} \tag{5.19}$$

$$\text{bias}_{(\text{YD.PX})} \triangleq \beta_{Y\sim D|P,X} - \beta_{Y\sim D|P,Z,X} = k \times \left(\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z,X}\right) \times \frac{\text{SD}(P^{\perp D,X})}{\text{SD}(D^{\perp P,X})} \tag{5.20}$$

$$\beta_{Y\sim D|P,Z,X} = \beta_{Y\sim D|P,X} - k \times \left(\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z,X}\right) \times \frac{\text{SD}(P^{\perp D,X})}{\text{SD}(D^{\perp P,X})} \tag{5.21}$$

Equations 5.20 and 5.21 contain the results. All components of these expressions except for $k \triangleq \frac{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{D\sim Z_{(\text{Y.DPX})}|P,X}}{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{P\sim Z_{(\text{Y.DPX})}|D,X}}$ and $\beta_{Y\sim P|D,Z,X}$ can be estimated from the data. We can use this for partial identification by considering plausible ranges of values for k and $\beta_{Y\sim P|D,Z,X}$. $\beta_{Y\sim P|D,Z,X}$ captures any direct causal relationship between P and Y. Since both P and Y are observed variables, their relationship conditional on D, Z, X should be able to be specified as the regression coefficient $\beta_{Y\sim P|D,Z,X}$ despite scale considerations. As with placebo outcomes, k captures the relative levels of confounding. Note that we can compare the level of confounding of the $P \rightarrow Y$ relationship from $\mathbf{Z}$ to the level of confounding of the $D \rightarrow Y$ relationship from $\mathbf{Z}$ by using the same linear combination, $Z_{(\text{Y.DPX})} = \mathbf{Z}\beta_{Y\sim Z|D,P,X}$. This is because in the placebo treatment approach we run a single regression with Y as the outcome that includes both D and P on the right-hand side. Since we assume that $\mathbf{Z}$ contains a rich enough set of unobserved variables to de-confound both $P \to Y$ and $D \to Y$, $Z_{(\text{Y.DPX})}$ is also sufficient for this. k is the ratio of the level of total confounding of $D \to Y$ from $D \leftarrow Z \to Y$ to the level of total confounding of $P \to Y$ from $P \leftarrow Z \to Y$, after re-scaling one of these biases to account for scale differences between P and D.[12]

Estimator and partial identification Equation 5.21 provides us with a moment estimator for $\beta_{Y\sim D|P,Z,X}$ in Equation 5.22.[13]

$$\hat{\beta}_{Y\sim D|P,Z,X} = \hat{\beta}_{Y\sim D|P,X} - k \times \left(\hat{\beta}_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z,X}\right) \times \frac{\|P^{\perp D,X}\|_2}{\|D^{\perp P,X}\|_2} \tag{5.22}$$

Given k and $\beta_{Y\sim P|D,Z,X}$, the estimator $\hat{\beta}_{Y\sim D|P,Z,X}$ from Equation 5.22 is consistent for $\beta_{Y\sim D|P,Z,X}$ and asymptotically normal. We discuss computing standard errors in Section 5.1.5. Using Equation 5.22 in a partial identification framework, we can reason about plausible ranges of values for $\beta_{Y\sim P|D,Z,X}$ and k to arrive at a range of plausible estimates of $\beta_{Y\sim D|P,Z,X}$. In the special case when there is only a single omitted variable, then k can also be written as $k = \frac{f_{D\sim Z|P,X}}{f_{P\sim Z|D,X}}$. This can be thought of as comparing the strength of the $Z \to D$ "arm" of the confounding path $D \leftarrow Z \to Y$ to the relative strength of the $Z \to P$ "arm" of the confounding path $P \leftarrow Z \to Y$ in Figure 5.1(b).

[12] Note that we do not cancel terms in $k = \frac{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X}\, f_{D\sim Z_{(\text{Y.DPX})}|P,X}}{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X}\, f_{P\sim Z_{(\text{Y.DPX})}|D,X}}$ since this would force us to reason about how $f_{D\sim Z_{(\text{Y.DPX})}|P,X}$ differs from $f_{D\sim \mathbf{Z}|P,X}$ and how $f_{P\sim Z_{(\text{Y.DPX})}|D,X}$ differs from $f_{P\sim \mathbf{Z}|D,X}$. It is more intuitive to reason about relative levels of bias, which does not require us to directly consider those differences.

[13] It is also possible to show that $\frac{\widehat{\text{SD}}(P^{\perp D,X})}{\widehat{\text{SD}}(D^{\perp P,X})} = \frac{\widehat{\text{SD}}(P^{\perp D,X})}{\widehat{\text{SD}}(D^{\perp P,X})} \times \frac{\widehat{\text{SD}}(Y^{\perp D,P,X})}{\widehat{\text{SD}}(Y^{\perp D,P,X})} = \frac{\text{se}(\hat{\beta}_{Y\sim D|P,X})}{\text{se}(\hat{\beta}_{Y\sim P|D,X})} \times \sqrt{\frac{\text{df}_D}{\text{df}_P}}$, which is an expression containing only commonly reported regression summary statistics from the short regression that is estimable from the observed data. $\text{se}(\hat{\beta}_{Y\sim D|P,X})$ and $\text{se}(\hat{\beta}_{Y\sim P|D,X})$ are the estimated standard errors from standard regression estimation of the short regression. df_Y and df_N are the degrees of freedom from the estimated short regression, these will be equal in this case.

5.1.4 Double placebos

Finally, we consider the setting in which we have two placebos. Suppose we are interested in $\beta_{Y\sim D|Z,X}$ from the "long" regression in Equation 5.1. However, Z is unobserved and so we must only consider estimating "short" regression in Equation 5.3. Further, suppose we are in a setting like Figure 5.2 - that is, we observe both a placebo outcome (N) and a placebo treatment (P). In such a "double placebo" setting, we can also develop a partial identification framework. Consider the following two sets of short and long regressions. Again, $\mathbf{Z}$ is an unobserved confounder. So we can only consider estimating the short regressions (i.e., Equations 5.23b and 5.23d). $Z_{(\mathrm{Y.DPX})} = \mathbf{Z}\beta_{Y\sim Z|D,P,X}$ is the linear combination that de-confounds the Y, D and the Y, P relationship. $Z_{(\mathrm{N.DPX})} = \mathbf{Z}\beta_{N\sim Z|D,P,X}$ is the linear combination that de-confounds the N, D and the N, P relationship. We choose $\mathbf{Z}$ to be a rich enough set of variables to allow for this interpretation.

$$Y = \beta_{Y\sim D|P,Z,X} D + \beta_{Y\sim P|D,Z,X} P + \mathbf{X}\beta_{Y\sim X|D,P,Z} + \mathbf{Z}\beta_{Y\sim Z|D,P,X} + \epsilon_{y,l} \tag{5.23a}$$

$$Y = \beta_{Y\sim D|P,X} D + \beta_{Y\sim P|D,X} P + \mathbf{X}\beta_{Y\sim X|D,P} + \epsilon_{y,s} \tag{5.23b}$$

$$N = \beta_{N\sim D|P,Z,X} D + \beta_{N\sim P|D,Z,X} P + \mathbf{X}\beta_{N\sim X|D,P,Z} + \mathbf{Z}\beta_{N\sim Z|D,P,X} + \epsilon_{n,l} \tag{5.23c}$$

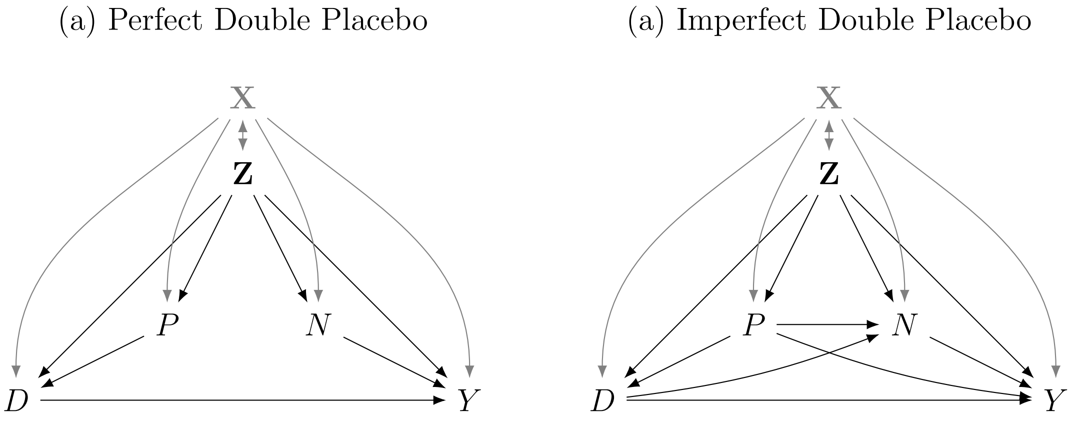
$$N = \beta_{N\sim D|P,X} D + \beta_{N\sim P|D,X} P + \mathbf{X}\beta_{N\sim X|D,P} + \epsilon_{n,s} \tag{5.23d}$$

In Appendix D.1, we detail a similar omitted variables bias based approach for this setting. The approach results in the expression for $\beta_{Y\sim D|P,Z_{(\mathrm{Y.DPX})},X} = \beta_{Y\sim D|P,Z,X}$ in Equation 5.24.

$$\beta_{Y\sim D|P,Z,X} = \beta_{Y\sim D|P,X} - k_{\left(\frac{\mathrm{YD}}{\mathrm{YP}}\right)} \times k_{\left(\frac{\mathrm{NP}}{\mathrm{ND}}\right)} \times \frac{(\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z,X})(\beta_{N\sim D|P,X} - \beta_{N\sim D|P,Z,X})}{(\beta_{N\sim P|D,X} - \beta_{N\sim P|D,Z,X})} \tag{5.24}$$

Equation 5.24 provides us with a moment estimator for $\beta_{Y\sim D|P,Z,X}$ in Equation 5.25.

Figure 5.2: A single causal graph with both a placebo outcome, N, and with a placebo treatment, P. D is treatment, Y is outcome, and $\mathbf{Z}$ contains unobserved confounders. $\mathbf{X}$ contains observed covariates.


(a) Perfect Double Placebo

(a) Imperfect Double Placebo

$$\hat{\beta}_{Y\sim D|P,Z,X} = \hat{\beta}_{Y\sim D|P,X} - k_{\left(\frac{\text{YD}}{\text{YP}}\right)} \times k_{\left(\frac{\text{NP}}{\text{ND}}\right)} \times \frac{(\hat{\beta}_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z,X})(\hat{\beta}_{N\sim D|P,X} - \beta_{N\sim D|P,Z,X})}{(\hat{\beta}_{N\sim P|D,X} - \beta_{N\sim P|D,Z,X})} \tag{5.25}$$

Given $k_{\left(\frac{\text{YD}}{\text{YP}}\right)}, k_{\left(\frac{\text{NP}}{\text{ND}}\right)}, \beta_{Y\sim P|D,Z,X}, \beta_{N\sim D|P,Z,X}$, and $\beta_{N\sim P|D,Z,X}$, the estimator $\hat{\beta}_{Y\sim D|P,Z,X}$ from Equation 5.25 is consistent for $\beta_{Y\sim D|P,Z,X}$ and asymptotically normal. We discuss computing standard errors in Section 5.1.5. Using Equation 5.25 in a partial identification framework, we can reason about plausible ranges of values for $k_{\left(\frac{\text{YD}}{\text{YP}}\right)}, k_{\left(\frac{\text{NP}}{\text{ND}}\right)}, \beta_{Y\sim P|D,Z,X}, \beta_{N\sim D|P,Z,X}$, and $\beta_{N\sim P|D,Z,X}$ to arrive at a range of plausible estimates of $\beta_{Y\sim D|P,Z,X}$.

In Figure 5.2(b), the partial identification parameters for double placebos capture paths as follows: $\beta_{Y\sim P|D,Z,X} = \beta_{Y\sim P|D,Z_{(\text{Y.DPX})},X}$ captures $P \rightarrow Y$ and $P \rightarrow N \rightarrow Y$ (the effect of P on Y conditional on D); $\beta_{N\sim D|P,Z,X} = \beta_{N\sim D|P,Z_{(\text{N.DPX})},X}$ captures $D \rightarrow N$; $\beta_{N\sim P|D,Z,X} = \beta_{N\sim P|D,Z_{(\text{N.DPX})},X}$ captures $P \rightarrow N$. We can disentangle the $P \rightarrow Y$ and $P \rightarrow N \rightarrow Y$ components of $\beta_{Y\sim P|D,Z,X}$ for Figure 5.2(b) by using the following omitted variable expression: $\beta_{Y\sim P|D,Z,X} = \beta_{Y\sim P|N,D,Z,X} + \beta_{Y\sim N|P,D,Z,X}\beta_{N\sim P|D,Z,X}$. $\beta_{Y\sim P|N,D,Z,X}$ captures $P \rightarrow Y$; $\beta_{Y\sim N|P,D,Z,X}$ captures $N \rightarrow Y$; $\beta_{N\sim P|D,Z,X}$ captures $P \rightarrow N$ and already a partial identification parameter. $k_{(\frac{\text{YD}}{\text{YP}})} \times k_{(\frac{\text{NP}}{\text{ND}})} = \frac{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{Z_{(\text{Y.DPX})}\sim D|P,X}}{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{Z_{(\text{Y.DPX})}\sim P|D,X}} \times \frac{R_{N\sim Z_{(\text{N.DPX})}|D,P,X} f_{Z_{(\text{N.DPX})}\sim P|D,X}}{R_{N\sim Z_{(\text{N.DPX})}|D,P,X} f_{Z_{(\text{N.DPX})}\sim D|P,X}}$. $k_{(\frac{\text{YD}}{\text{YP}})}$ captures the scaled ratio of the level of confounding of the Y, D relationship (conditional on P) to the level of confounding of the Y, P relationship (conditional on D), after re-scaling one of these biases, or the ratio of bias factors. $k_{(\frac{\text{NP}}{\text{ND}})}$ captures the scaled ratio of the level of confounding of the N, P relationship (conditional on D) to the level of confounding of the N, D relationship (conditional on P), after re-scaling one of these biases, or the ratio of bias factors.

This type of expression can be used in a wide variety of settings. Though the interpretation of the partial identification parameters and of $\beta_{Y\sim D|P,Z,X}$ will depend on the underlying causal model being considered. See Section 5.2 for more details. With two imperfect placebos, users have a choice between the "single placebo" approaches and the "double placebo" approach. If we are in a setting like in Figure 5.2(a), where "perfect" placebos are available, then $\beta_{Y\sim P|D,Z,X}$, $\beta_{N\sim D|P,Z,X}$, and $\beta_{N\sim P|D,Z,X}$ will all equal zero. If it is reasonable to assume that either we have a single unobserved variable, $Z_{(\text{Y.DPX})} = Z_{(\text{N.DPX})}$, or $R_{Z_{(\text{Y.DPX})}\sim D|P,X} = R_{Z_{(\text{N.DPX})}\sim D|P,X}$ and $R_{Z_{(\text{Y.DPX})}\sim P|D,X} = R_{Z_{(\text{N.DPX})}\sim P|D,X}$, then $k_{(\frac{\text{YD}}{\text{YP}})} \times k_{(\frac{\text{NP}}{\text{ND}})} = 1$.[14] If we are in both of these settings at the same time, then we can point identify $\beta_{Y\sim D|Z,X}$ using observed data. This is similar to the proximal causal inference point identification results (Miao et al., 2020). See Liu et al. (2022) for an application of a double placebo approach.

5.1.5 Standard Errors

For fixed values for the partial identification parameters, the moment estimators in Equations 5.16, 5.22, and 5.25 are consistent for the target regression coefficient on the treatment from the "long" regression and are asymptotically normal. We can use the non-parametric bootstrap to estimate standard errors. See Zhang and Ding (2022); Freidling and Zhao (2023)

[14] If these relationships approximately hold, we could also consider a narrow range of values for $k_{(\frac{\text{YD}}{\text{YP}})} \times k_{(\frac{\text{NP}}{\text{ND}})}$ close to 1.

for similar partial identification approaches that rely on the non-parametric bootstrap for standard errors.

5.2 A taxonomy of causal graphs

How do we know if we have a useful placebo? How do we know if it is better to treat a placebo as a placebo outcome or a placebo treatment? Are there settings in which the above approaches should not be used? Are there settings in which the proposed methods can be used for problems other than unobserved confounding? We will address these questions by developing a taxonomy of causal graphs when a single placebo is available and when two are available. For introductions to graphical causal models see Pearl (2009), Richardson and Robins (2013a), Matthay and Glymour (2020), and Cinelli et al. (2022). We re-emphazise the Necker cube quality of placebos. Two causal graphs, each with a placebo, that have the same structure may invite different interpretations for whether the placebo is a placebo treatment or outcome. These choices are context specific. Sometimes the placebo may feel more like an outcome, while in other contexts the placebo feels more like a treatment.

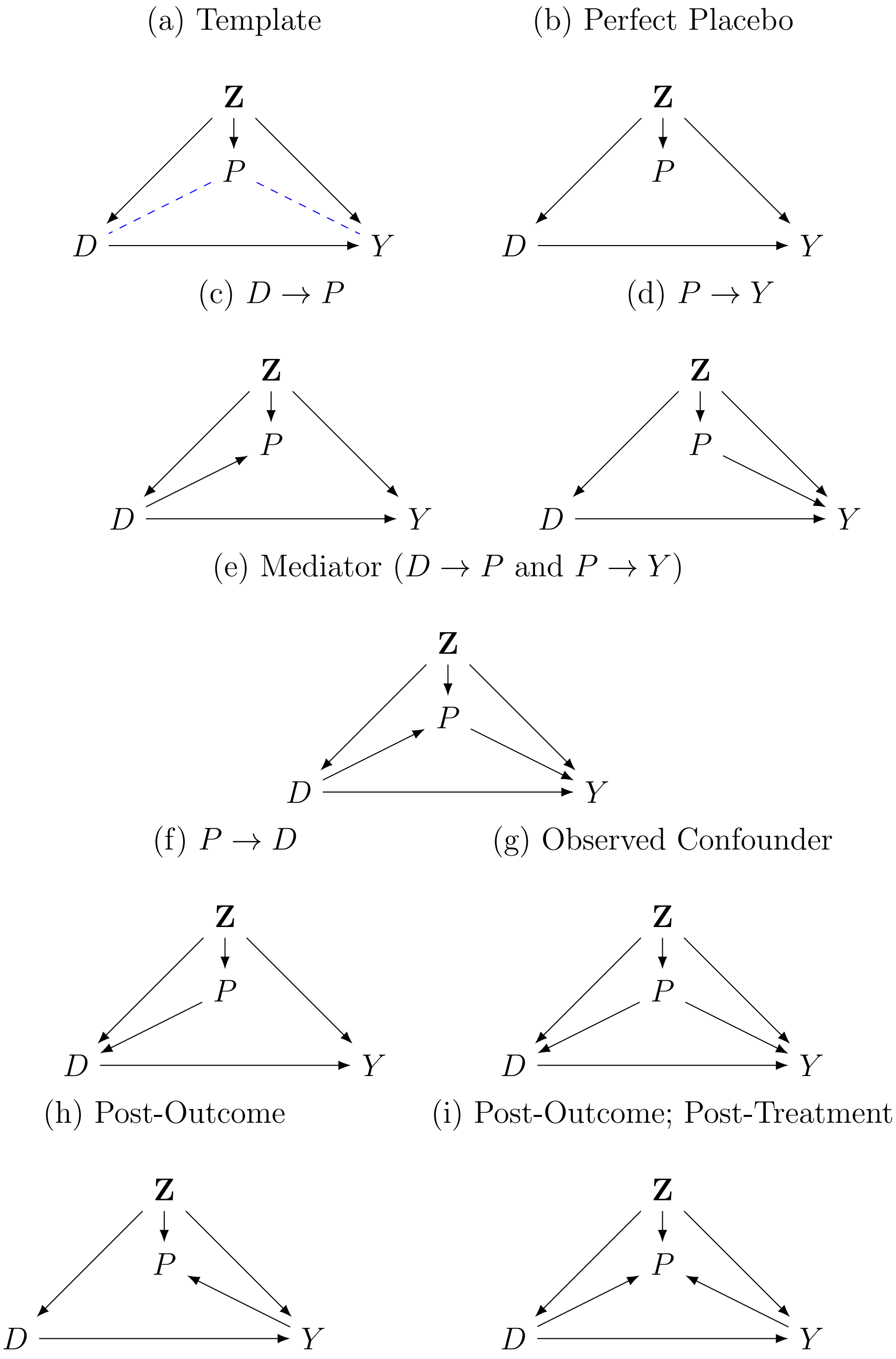
5.2.1 Single Placebos

We start by listing all possible causal graphs when a single placebo is observed in Figure 5.3. We assume that all relevant unobserved common causes are bundled into the vector of unobserved variables $\mathbf{Z}$. Observed covariates $\mathbf{X}$ are omitted from the graphs for clarity. We refer to the placebo as P. Figure 5.3(a) contains a template from which the rest of the graphs in Figure 5.3 are derived. The two dashed edges can point in either direction or not exist.[15] Different approaches should be used when dealing with each causal structure.[16]

[15] There are $3^2 - 1 = 8$ possible graphs, excluding those with cycles.

[16] The approaches presented in Section 5.1.2 and 5.1.3 hold for any causal graph. However, the interpretation of the partial identification parameters depend on the causal graph. We, therefore, provide guidance on which approaches we suggest for each causal graph. Assuming a causal graph (or set of causal graphs) that are

Figure 5.3: Single placebo DAGs.



(a) Template

(b) Perfect Placebo

(c) $D \rightarrow P$

(d) $P \rightarrow Y$

(e) Mediator ($D \rightarrow P$ and $P \rightarrow Y$)

(f) $P \rightarrow D$

(g) Observed Confounder

(h) Post-Outcome

(i) Post-Outcome; Post-Treatment

plausible in any application is a pre-requisite to making any credible causal claims. This section includes some additional approaches not yet discussed. Details of these can be found in the Appendix.

Placebos Figure 5.3(b) is the simplest setting in which P has no direct relationship with either the treatment, D, or the outcome, Y. That is, P is a "perfect" placebo. There are two options for how we might proceed in this case. We can look at P as a placebo outcome and follow the approach laid out in Section 5.1.2, or we can look at P as a placebo treatment and follow the approach laid out in Section 5.1.3. In either case, since P is a perfect placebo, we will only have one partial identification parameter that captures the relative levels of confounding.

Figure 5.3(c) has the addition of $D \to P$. Since the placebo is post-treatment in this setting, it is likely that we will want to treat the placebo as an imperfect placebo outcome. Therefore, we can follow the approach laid out in Section 5.1.2 [17] Figure 5.3(d) has $P \to Y$. Here there are also two options. The first would be to treat P as an imperfect placebo treatment and follow the approach laid out in Section 5.1.3. The second would be to view P as a placebo outcome, as in the case that P is a pre-treatment version of the outcome. In this setting, we can employ the approach laid out in Section 5.1.2 or Equation 5.26. The key in using Equation 5.26 is the inclusion of P in the regression of Y on D.[18]

[17] In Figure 5.3(c), the approach from Section 5.1.3 will not work well. The regression $Y \sim D + P$ will run into the problem that $\beta_{Y\sim D|P} = \beta_{Y\sim D|P,Z} + \beta_{Y\sim Z|D,P}\beta_{Z\sim D|P}$ and $\beta_{Z\sim D|P}$ captures both $Z \to D$ and $Z \to P \leftarrow D$. This means that a parameter that we will need to reason about will capture the effect of conditioning on a collider.

[18] Note that Chabé-Ferret (2017), Ham and Miratrix (2022), and others warn that conditioning on pre-treatment outcomes in a difference-in-differences type approach can create amplify bias under certain circumstances. However, since we take a partial identification approach, this should not present any issues for using Equation 5.26.

$$
\begin{aligned}
&\text{bias}_{(\text{YD.PX})} = \beta_{Y\sim D|P,X} - \beta_{Y\sim D|P,Z_{(\text{Y.DPX})},X} \\
&= R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{D\sim Z_{(\text{Y.DPX})}|P,X} (\text{SD}(Y^{\perp D,P,X})/\text{SD}(D^{\perp P,X})) \\
&\text{bias}_{(\text{PD.X})} = \beta_{P\sim D|X} - \beta_{P\sim D|Z_{(\text{P.DX})},X} \\
&= R_{P\sim Z_{(\text{P.DX})}|D,X} f_{D\sim Z_{(\text{P.DX})}|X} (\text{SD}(P^{\perp D,X})/\text{SD}(D^{\perp X})) \\
&\text{bias}_{(\text{YD.PX})} = m \times \text{bias}_{(\text{PD.X})} \\
&\implies \beta_{Y\sim D|P,Z_{(\text{Y.DPX})},X} = \beta_{Y\sim D|P,X} - m \times (\beta_{P\sim D|X} - \beta_{P\sim D|Z_{(\text{P.DX})},X}) \qquad (5.26) \\
&m = \frac{\text{bias}_{(\text{YD.PX})}}{\text{bias}_{(\text{PD.X})}} = \frac{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{D\sim Z_{(\text{Y.DPX})}|P,X}}{R_{P\sim Z_{(\text{P.DX})}|D,X} f_{D\sim Z_{(\text{P.DX})}|X}} \times \frac{\text{SD}(Y^{\perp D,P,X})}{\text{SD}(D^{\perp P,X})} \times \frac{\text{SD}(D^{\perp X})}{\text{SD}(P^{\perp D,X})} \\
&= k \times \frac{\text{SD}(Y^{\perp D,P,X})}{\text{SD}(D^{\perp P,X})} \times \frac{\text{SD}(D^{\perp X})}{\text{SD}(P^{\perp D,X})} \\
&\therefore \beta_{Y\sim D|P,Z_{(\text{Y.DPX})},X} \\
&= \beta_{Y\sim D|P,X} - k \times (\beta_{P\sim D|X} - \beta_{P\sim D|Z_{(\text{P.DX})},X}) \times \frac{\text{SD}(Y^{\perp D,P,X})}{\text{SD}(D^{\perp P,X})} \times \frac{\text{SD}(D^{\perp X})}{\text{SD}(P^{\perp D,X})}
\end{aligned}
$$

In Equation 5.26, $k \triangleq \frac{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{D\sim Z_{(\text{Y.DPX})}|P,X}}{R_{P\sim Z_{(\text{P.DX})}|D,X} f_{D\sim Z_{(\text{P.DX})}|X}}$ is the ratio of the level of confounding D and Y, conditional on P, to the level of confounding D and P, after re-scaling one of these biases to account for scale differences. Since we included P in the regression of Y on D, we see that the confounding of D and Y does not include the $Z \to P \to Y$ path, possibly making comparison with the confounding of D and P easier. $\beta_{P\sim D|Z_{(\text{P.DX})},X} = \beta_{P\sim D|Z,X}$ measures the causal effect of D on P, which based on Figure 5.3(d) should be zero.

Mediators Figure 5.3(e) has $D \to P$ and $P \to Y$, making P a mediator between D and Y. In this case there is more than one option. First, Zhang and Ding (2022) provide guidance on how to take an omitted variables bias approach to partial identification of the direct effect and indirect effect. The direct effect approach deals with the fact that conditioning on P opens the collider path $D \to P \leftarrow Z$ and involves specifying partial identification parameters for each of the $Z \to D$, $Z \to P$, and $Z \to Y$ relationships. Suppose we want to consider the total effect. We could use the approach outlined in Cinelli and Hazlett (2020), but this approach for the total effect applied to a setting with a mediator requires that we reason about $R^2_{Y\sim Z|D}$ which captures the $Z \to Y$ and $Z \to P \to Y$ relationships. A wrinkle is that $P \to Y$ is a component of the total effect we want to identify. An alternative would be to follow the approach laid out in Section 5.1.2, where the mediator is treated as a placebo outcome. Here there is again the wrinkle that we will be required to reason about $D \to P$ (above we refer to the placebo outcomes as N), which is a component of the total effect we want to identify. We could also follow the approach in Equation 5.27. This does something similar but where we treat the mediator as a placebo treatment. The wrinkle remains that we need to reason about $P \to Y$, a component of the total effect we want to identify.

$$\begin{aligned}
&\text{bias}_{(\text{YD.X})} = \beta_{Y\sim D|X} - \beta_{Y\sim D|Z_{(\text{Y.DX})},X} \\
&= R_{Y\sim Z_{(\text{Y.DX})}|D,X} f_{D\sim Z_{(\text{Y.DX})}|X} (\text{SD}(Y^{\perp D,X})/\text{SD}(D^{\perp X})) \\
&\text{bias}_{(\text{YP.DX})} = \beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\text{Y.DPX})},X} \\
&= R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{P\sim Z_{(\text{Y.DPX})}|D,X} (\text{SD}(Y^{\perp D,P,X})/\text{SD}(P^{\perp D,X})) \\
&\text{bias}_{(\text{YD.X})} = m \times \text{bias}_{(\text{YP.DX})} \\
&\implies \beta_{Y\sim D|Z_{(\text{Y.DX})},X} = \beta_{Y\sim D|X} - m \times (\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\text{Y.DPX})},X}) \qquad (5.27) \\
&m = \frac{\text{bias}_{(\text{YD.X})}}{\text{bias}_{(\text{YP.DX})}} = \frac{R_{Y\sim Z_{(\text{Y.DX})}|D,X} f_{D\sim Z_{(\text{Y.DX})}|X}}{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{P\sim Z_{(\text{Y.DPX})}|D,X}} \frac{\text{SD}(P^{\perp D,X})}{\text{SD}(D^{\perp X})} \frac{\text{SD}(Y^{\perp D,X})}{\text{SD}(Y^{\perp D,P,X})} \\
&= k \times \frac{\text{SD}(P^{\perp D,X})}{\text{SD}(D^{\perp X})} \frac{\text{SD}(Y^{\perp D,X})}{\text{SD}(Y^{\perp D,P,X})} \\
&\therefore \beta_{Y\sim D|Z_{(\text{Y.DX})},X} \\
&= \beta_{Y\sim D|X} - k \times (\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\text{Y.DPX})},X}) \times \frac{\text{SD}(P^{\perp D,X})}{\text{SD}(D^{\perp X})} \frac{\text{SD}(Y^{\perp D,X})}{\text{SD}(Y^{\perp D,P,X})}
\end{aligned}$$

In Equation 5.27, $k \triangleq \frac{R_{Y\sim Z_{(\text{Y.DX})}|D,X} f_{D\sim Z_{(\text{Y.DX})}|X}}{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X} f_{P\sim Z_{(\text{Y.DPX})}|D,X}}$ is the ratio of the level of confounding

D and Y to the level of confounding P and Y, conditional on D, after re-scaling one of these biases to account for scale differences. $\beta_{Y\sim P|D,Z_{\text{(Y.DPX)}},X} = \beta_{Y\sim P|D,Z,X}$ measures the causal effect of P on Y.

Observed Confounders Figure 5.3(f) has $P \rightarrow D$. This is similar to an observed confounder. Figure 5.3(g) has $P \rightarrow D$ and $P \rightarrow Y$, making P an observed confounder. In both of these cases, there are two options. First, we could follow the approach laid out in Cinelli and Hazlett (2020), where P is an observed covariate. Second, we could follow the approach in Equation 5.28. The key here is to include P in the regression of Y on D and to also consider the regression of D on P.

$$
\begin{aligned}
&\text{bias}_{\text{(YD.PX)}} = \beta_{Y\sim D|P,X} - \beta_{Y\sim D|P,Z_{\text{(Y.DPX)}},X} \\
&= R_{Y\sim Z_{\text{(Y.DPX)}}|D,P,X} f_{D\sim Z_{\text{(Y.DPX)}}|P,X}(\text{SD}(Y^{\perp D,P,X})/\text{SD}(D^{\perp P,X})) \\
&\text{bias}_{\text{(DP.X)}} = \beta_{D\sim P|X} - \beta_{D\sim P|Z_{\text{(D.PX)}},X} \\
&= R_{D\sim Z_{\text{(D.PX)}}|P,X} f_{P\sim Z_{\text{(D.PX)}}|X}(\text{SD}(D^{\perp P,X})/\text{SD}(P^{\perp X})) \\
&\text{bias}_{\text{(YD.PX)}} = m \times \text{bias}_{\text{(DP.X)}} \\
&\implies \beta_{Y\sim D|P,Z_{\text{(Y.DPX)}},X} = \beta_{Y\sim D|P,X} - m \times (\beta_{D\sim P|X} - \beta_{D\sim P|Z_{\text{(D.PX)}},X}) \\
&m = \frac{\text{bias}_{\text{(YD.PX)}}}{\text{bias}_{\text{(DP.X)}}} = \frac{R_{Y\sim Z_{\text{(Y.DPX)}}|D,P,X} f_{D\sim Z_{\text{(Y.DPX)}}|P,X}}{R_{D\sim Z_{\text{(D.PX)}}|P,X} f_{P\sim Z_{\text{(D.PX)}}|X}} \times \frac{\text{SD}(Y^{\perp D,P,X})}{\text{SD}(D^{\perp P,X})} \times \frac{\text{SD}(P^{\perp X})}{\text{SD}(D^{\perp P,X})} \\
&= k \times \frac{\text{SD}(Y^{\perp D,P,X})}{\text{SD}(D^{\perp P,X})} \times \frac{\text{SD}(P^{\perp X})}{\text{SD}(D^{\perp P,X})} \\
&\therefore \beta_{Y\sim D|P,Z_{\text{(Y.DPX)}},X} \\
&= \beta_{Y\sim D|P,X} - k \times (\beta_{D\sim P|X} - \beta_{D\sim P|Z_{\text{(D.PX)}},X}) \times \frac{\text{SD}(Y^{\perp D,P,X})}{\text{SD}(D^{\perp P,X})} \times \frac{\text{SD}(P^{\perp X})}{\text{SD}(D^{\perp P,X})}
\end{aligned}
\tag{5.28}
$$

In Equation 5.28, $k \triangleq \frac{R_{Y\sim Z_{\text{(Y.DPX)}}|D,P,X} f_{D\sim Z_{\text{(Y.DPX)}}|P,X}}{R_{D\sim Z_{\text{(D.PX)}}|P,X} f_{P\sim Z_{\text{(D.PX)}}|X}}$ is the ratio of the level of confounding D and Y, conditional on P, to the level of confounding D and P, after re-scaling one of these biases to account for scale differences. Since we included P in the regression of Y on D, we see that the confounding of D and Y does not include the paths that run through P, making comparison with the confounding of D and P easier. $\beta_{D\sim P|Z_{(\mathrm{D.PX})},X} = \beta_{D\sim P|Z,X}$ measures the causal effect of P on D.

Post-outcome Figure 5.3(h) has $Y \to P$. Figure 5.3(i) has $Y \to P$ and $D \to P$. In both cases, P is a post-outcome variable. In both of these cases, there are two options. First, we could follow the approach laid out in Cinelli and Hazlett (2020), where we ignore P. Second, we could follow the approach in Equation 5.29. The key is to consider the regression of P on Y and D.

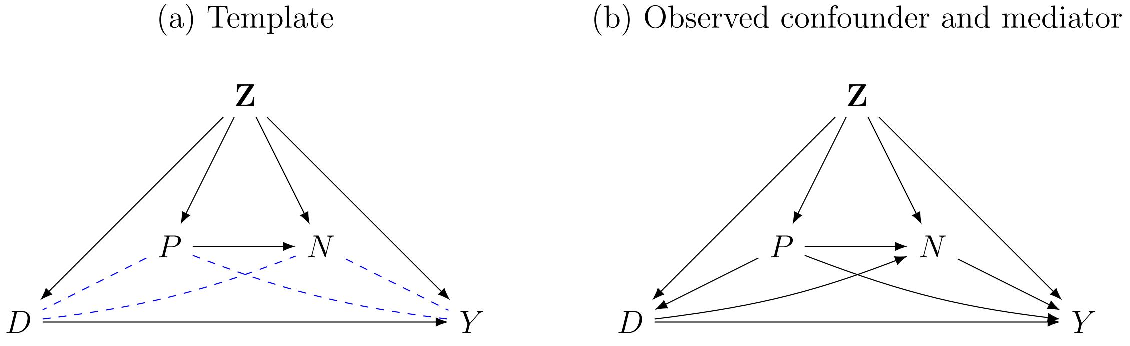
$$\begin{aligned}
&\mathrm{bias}_{(\mathrm{YD.X})} = \beta_{Y\sim D|X} - \beta_{Y\sim D|Z_{(\mathrm{Y.DX})},X} \\
&= R_{Y\sim Z_{(\mathrm{Y.DX})}|D,X}\, f_{D\sim Z_{(\mathrm{Y.DX})}|X} (\mathrm{SD}(Y^{\perp D,X})/\mathrm{SD}(D^{\perp X})) \\
&\mathrm{bias}_{(\mathrm{PY.DX})} = \beta_{P\sim Y|D,X} - \beta_{P\sim Y|D,Z_{(\mathrm{P.YDX})},X} \\
&= R_{P\sim Z_{(\mathrm{P.YDX})}|Y,D,X}\, f_{Y\sim Z_{(\mathrm{P.YDX})}|D,X} (\mathrm{SD}(P^{\perp Y,D,X})/\mathrm{SD}(Y^{\perp D,X})) \\
&\mathrm{bias}_{(\mathrm{YD.X})} = m \times \mathrm{bias}_{(\mathrm{PY.DX})} \\
&\implies \beta_{Y\sim D|Z_{(\mathrm{Y.DX})},X} = \beta_{Y\sim D|X} - m \times (\beta_{P\sim Y|D,X} - \beta_{P\sim Y|D,Z_{(\mathrm{P.YDX})},X}) \\
&m = \frac{\mathrm{bias}_{(\mathrm{YD.X})}}{\mathrm{bias}_{(\mathrm{PY.DX})}} = \frac{R_{Y\sim Z_{(\mathrm{Y.DX})}|D,X}\, f_{D\sim Z_{(\mathrm{Y.DX})}|X}}{R_{P\sim Z_{(\mathrm{P.YDX})}|Y,D,X}\, f_{Y\sim Z_{(\mathrm{P.YDX})}|D,X}} \times \frac{\mathrm{SD}(Y^{\perp D,P,X})}{\mathrm{SD}(D^{\perp P,X})} \times \frac{\mathrm{SD}(P^{\perp X})}{\mathrm{SD}(D^{\perp P,X})} \\
&= k \times \frac{\mathrm{SD}(Y^{\perp D,X})}{\mathrm{SD}(D^{\perp X})} \times \frac{\mathrm{SD}(Y^{\perp D,X})}{\mathrm{SD}(P^{\perp Y,D,X})} \\
&\therefore \beta_{Y\sim D|Z_{(\mathrm{Y.DX})},X} \\
&= \beta_{Y\sim D|X} - k \times (\beta_{P\sim Y|D,X} - \beta_{P\sim Y|D,Z_{(\mathrm{P.YDX})},X}) \times \frac{\mathrm{SD}(Y^{\perp D,X})}{\mathrm{SD}(D^{\perp X})} \times \frac{\mathrm{SD}(Y^{\perp D,X})}{\mathrm{SD}(P^{\perp Y,D,X})}
\end{aligned} \tag{5.29}$$

In Equation 5.29, $k \triangleq \frac{R_{Y\sim Z_{(\mathrm{Y.DX})}|D,X}\, f_{D\sim Z_{(\mathrm{Y.DX})}|X}}{R_{P\sim Z_{(\mathrm{P.YDX})}|Y,D,X}\, f_{Y\sim Z_{(\mathrm{P.YDX})}|D,X}}$ is the ratio of the level of confounding of D and Y to the level of confounding of Y and P, after re-scaling one of these biases to account for scale differences. $\beta_{P\sim Y|D,Z_{(\mathrm{P.YDX})},X} = \beta_{P\sim Y|D,Z,X}$ measures the causal effect of Y on P.

5.2.2 Double Placebos

In the double placebo case, there are many more possible graphs. Again, we assume that all relevant unobserved common causes are bundled into the vector of unobserved variables $\mathbf{Z}$. Observed covariates $\mathbf{X}$ are omitted from the graphs for clarity. All the dashed edges in Figure 5.4(a) could have either direction or not exist[19] (excluding combinations of edge directions that lead to cycles). However, we limit our analysis to the specific case shown in Figure 5.4(b) and detailed in Section 5.1.4. This may cover many practical settings and researchers interested in alternative graphs can reason through the implications of their favored graph and its implications for interpreting the partial identification parameters based on Section 5.1.4. There may be other partial identification approaches available in the double placebo case but we omit an exploration of these for brevity.

Figure 5.4: Double placebo causal graph.


(a) Template

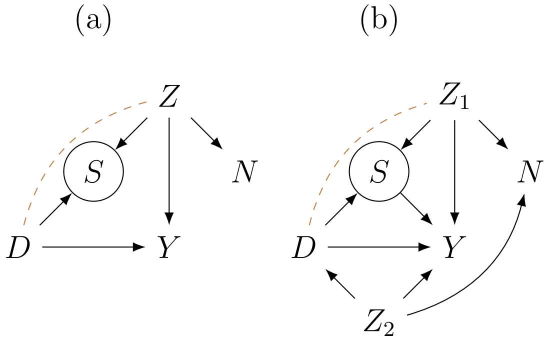
(b) Observed confounder and mediator

[19] There are $3^4 = 81$ possible graphs, assuming that users know which of the two placebos is causally prior to the other; but the graphs with cycles should be excluded. Making the conservative assumption that the dashed edges in Figure 5.4(a) do exist, then there are $2^4 = 16$ possible fully connected graphs; again, the graphs with cycles should be excluded.

5.2.3 Sample selection and other applications

Arnold et al. (2016) suggests that placebos can be used to address sample selection and measurement bias in addition to unobserved confounding. This might look like Figure 5.5. Take Figure 5.5(a) as an example. Here selection of units into the study sample is indicated by a special node S. Since we are limited to the study sample, we are implicitly conditioning on S, which opens a non-causal path from D to Y. However, if sample selection also opens a similar non-causal path from D to N, then we may be able to use information about the placebo outcome N to inform how sample selection biases the estimate of the effect of D on Y. See Chapter 2 for a discussion of causal graphs and sample selection.

Figure 5.5: Example causal graphs with a placebo outcome, N. D is treatment; Y is outcome; Z contains unobserved confounders; S represents sample selection. Additional observed covariates are omitted for simplicity.



5.3 Difference in differences as special case

In this section, we explore how the approaches put forth in this paper correspond to various relaxations to the standard assumptions of difference in differences identification strategies. Let's start by reviewing the typical difference in differences (DID) setup and identification strategy. Suppose we have panel data for two periods for a cohort of units. We observe the outcome of interest in period 1, which we will label N, as well as the outcome of interest in period 2, which we will call Y. In period 1, no units are treated but in period 2 some units are treated. We indicate the group of units that is treated with $G = 1$. We indicate actual treatment with $D = 1$. We therefore have that in period 1 $D = 0$ for all units and that in period 2 $G = D$. For now, we assume that we do not need to adjust for any covariates. The standard DID identification strategy identifies the ATT for Y (ATT_Y) in the following way. Note that Y_d and N_d are the potential outcomes when D is set, perhaps counterfactually, to $D = d$ with $d \in \{0, 1\}$. We start by revisiting a standard decomposition of the simple difference in means (DIM) between the treated and control groups for Y.

$$\begin{aligned}
\text{DIM}_Y &= \mathbb{E}[Y|G=1] - \mathbb{E}[Y|G=0] \\
&= \mathbb{E}[Y_1|G=1] - \mathbb{E}[Y_0|G=0] \text{ by consistency and since } D = G \text{ in period 2} \\
&= \underbrace{\mathbb{E}[Y_1|G=1] - \mathbb{E}[Y_0|G=1]}_{\text{ATT}_Y} + \underbrace{\mathbb{E}[Y_0|G=1] - \mathbb{E}[Y_0|G=0]}_{\text{Bias}_Y} \\
&= \text{ATT}_Y + \text{Bias}_Y
\end{aligned} \tag{5.30}$$

We make a "parallel trends" assumption that the trend in the mean control potential outcomes for the treated group is the same as the trend for the control group (i.e., $\mathbb{E}[Y_0|G = 1] - \mathbb{E}[N_0|G = 1] = \mathbb{E}[Y_0|G = 0] - \mathbb{E}[N_0|G = 0]$). The parallel trends assumption can be re-ordered into an "equiconfounding" assumption that the mean baseline difference between the treated and control groups in period 2 is the same as the mean baseline difference between the treated and control groups in period 1. The equivalence of parallel trends and equiconfounding is discussed in Sofer et al. (2016).

$$\begin{aligned}
&\text{Parallel Trends} \triangleq \underbrace{\mathbb{E}[Y_0|G=1] - \mathbb{E}[N_0|G=1]}_{\text{Trend}_{G=1}} = \underbrace{\mathbb{E}[Y_0|G=0] - \mathbb{E}[N_0|G=0]}_{\text{Trend}_{G=0}} \\
&\iff \underbrace{\mathbb{E}[Y_0|G=1] - \mathbb{E}[Y_0|G=0]}_{\text{Bias}_Y} = \underbrace{\mathbb{E}[N_0|G=1] - \mathbb{E}[N_0|G=0]}_{\text{Bias}_N} \triangleq \text{Equiconfounding}
\end{aligned} \tag{5.31}$$

Now, we recall our constraint that N is a pre-treatment version of the outcome. This means that the treatment cannot have an effect on the pre-treatment outcome (and that no units are treated in period 1). So $N_1 = N_0 = N$ for both the treated and control groups. So we see that $\text{DIM}_N = \text{ATT}_N + \text{Bias}_N = \text{Bias}_N$.

$$\begin{aligned}
\text{DIM}_N &= \mathbb{E}[N|G=1] - \mathbb{E}[N|G=0] \\
&= \mathbb{E}[N_1|G=1] - \mathbb{E}[N_0|G=0] \\
&= \mathbb{E}[N_1|G=1] - \mathbb{E}[N_0|G=1] + \mathbb{E}[N_0|G=1] - \mathbb{E}[N_0|G=0] \\
&= \underbrace{\mathbb{E}[N_1 - N_0|G=1]}_{\text{ATT}_N} + \underbrace{\mathbb{E}[N_0|G=1] - \mathbb{E}[N_0|G=0]}_{\text{Bias}_N} \\
&= \mathbb{E}[N - N|G=1] + \text{Bias}_N = \text{Bias}_N
\end{aligned} \tag{5.32}$$

Therefore, we have that $\text{Bias}_Y = \text{Bias}_N = \text{DIM}_N$ and that $\text{DIM}_Y = \text{ATT}_Y + \text{Bias}_Y$. Plugging into the later, we have $\text{DIM}_Y = \text{ATT}_Y + \text{DIM}_N$, which can be rearranged to yield the standard difference in differences identification result, $\text{ATT}_Y = \text{DIM}_Y - \text{DIM}_N$, which makes clear where the DID name comes from.

Parallel Trends and equiconfounding While it is common to make parallel trends assumptions, equiconfounding is mathematically equivalent and is not any less intuitive to consider. equiconfounding assumes that the difference in mean control potential outcomes for the treated and control groups in period 2 is the same as the difference for the treated and control groups in period 1. This means that, on average, the way that the treatment and control groups differ for the outcome in period 1 can be used as a stand in for the way that the treatment and control groups differ for the outcome in period 2. Researchers should be able to consider this and possible violations to it just as easily as the parallel trends assumptions and its violations. In what follows, we will stick to equiconfounding type assumptions, rather than parallel trends assumptions since they have a more direct connection to the bias that we are trying to overcome, namely Bias_Y. But every equiconfounding type assumption we might consider has analogous parallel trends assumption, as we will show.

Relaxing Assumptions It is easy to see that we can relax the assumption that there is no effect of D on N by allowing ATT_N to be non-zero. Further, we can relax the assumption that Bias_Y exactly equals Bias_N by recognizing that, for some $m \in \mathbb{R}$, $\text{Bias}_Y = m \times \text{Bias}_N$. Together, these relaxations give us the expression for ATT_Y in Equation 5.33.

$$\begin{aligned} \text{ATT}_Y &= \text{DIM}_Y - \text{Bias}_Y \\ &= \text{DIM}_Y - m \times \text{Bias}_N \\ &= \text{DIM}_Y - m \times (\text{DIM}_N - \text{ATT}_N) \end{aligned} \tag{5.33}$$

That is, we can express ATT_Y as a difference between DIM_Y and an "adjusted" version of DIM_N, where the adjustment is parameterized by $m = \frac{\text{Bias}_Y}{\text{Bias}_N}$ and ATT_N. The statement $\text{Bias}_Y = m \times \text{Bias}_N$ is similar to the equiconfounding assumption from the standard DID approach. And the expression $\text{ATT}_Y = \text{DIM}_Y - m \times (\text{DIM}_N - \text{ATT}_N)$ is similar to a difference in differences type identification result. So we are not too far from familiar results, but this approach would apply to more general placebo outcomes and settings outside the standard DID setting. Equation 5.33 could be used for partial identification in cases with binary treatments and a placebo outcome. Researchers would estimate DIM_Y and DIM_N and then reason about plausible values for m and ATT_N to partially identify ATT_Y.[20] But how does

[20] We suspect there are more settings in which multiple applications of the decomposition DIM = ATT+Bias could yield useful expressions for causal effects with binary treatments, similar to how Section 5.1 involves multiple applications of the omitted variables bias formula. For example, another possibility is for a binary placebo treatment P with actual binary treatment D and outcome Y. We may have that

$$\underbrace{\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]}_{\text{DIM}_{Y\sim D}} = \underbrace{\mathbb{E}[Y_{D=1}|D=1] - \mathbb{E}[Y_{D=0}|D=1]}_{\text{ATT}_{Y\sim D}} + \underbrace{\mathbb{E}[Y_{D=0}|D=1] - \mathbb{E}[Y_{D=0}|D=0]}_{\text{Bias}_{Y\sim D}}$$

$$\underbrace{\mathbb{E}[Y|P=1] - \mathbb{E}[Y|P=0]}_{\text{DIM}_{Y\sim P}} = \underbrace{\mathbb{E}[Y_{P=1}|P=1] - \mathbb{E}[Y_{P=0}|P=1]}_{\text{ATT}_{Y\sim P}} + \underbrace{\mathbb{E}[Y_{P=0}|P=1] - \mathbb{E}[Y_{P=0}|P=0]}_{\text{Bias}_{Y\sim P}}$$

There is some m such that $\text{Bias}_{Y\sim D} = m \times \text{Bias}_{Y\sim P}$. So we can write $\text{ATT}_{Y\sim D} = \text{DIM}_{Y\sim D} - m \times (\text{DIM}_{Y\sim P} - \text{ATT}_{Y\sim P})$. We could then partially identify $\text{ATT}_{Y\sim D}$ by estimating $\text{DIM}_{Y\sim D}$ and $\text{DIM}_{Y\sim P}$ and reasoning about plausible values for $m \triangleq \frac{\text{Bias}_{Y\sim D}}{\text{Bias}_{Y\sim P}}$ and $\text{ATT}_{Y\sim P}$.

this all relate to the regression-based results we discuss in Section 5.1?

Connecting to Regression Since G is binary, a regression of Y or N on G will give the difference in means DIM_Y or DIM_N, respectively. Additionally, we can use traditional omitted variable bias analysis to show that,

$$\begin{aligned}\text{DIM}_Y &= \beta_{Y\sim G} = \beta_{Y\sim G|Z_{(\text{Y.G})}} + \beta_{Y\sim Z_{(\text{Y.G})}|G}\beta_{Z_{(\text{Y.G})}\sim G} \\ \text{DIM}_N &= \beta_{N\sim G} = \beta_{N\sim G|Z_{(\text{N.G})}} + \beta_{N\sim Z_{(\text{N.G})}|G}\beta_{Z_{(\text{N.G})}\sim G}\end{aligned} \tag{5.34}$$

where $Z_{(\text{Y.D})}$ and $Z_{(\text{N.D})}$ are linear combinations of a vector $\mathbf{Z}$ of all the time-varying and non-time-varying confounders of the G, Y and G, N relationships, like in previous sections. $\beta_{Y\sim G}$ is the regression coefficient on G from the regression $Y = \beta_{Y\sim G}G + \epsilon_1$. $\beta_{Y\sim G|Z} = \beta_{Y\sim G|Z_{(\text{Y.G})}}$ is the regression coefficient on G from the regression $Y = \beta_{Y\sim G|Z}G + \mathbf{Z}\beta_{Y\sim Z|G} + \epsilon_2$. All the other regression coefficients are defined similarly. We can also see that Equation 5.35 holds.

$$\text{ATT}_Y + \text{Bias}_Y = \text{DIM}_Y = \beta_{Y\sim G} = \beta_{Y\sim G|Z} + \beta_{Y\sim Z_{(\text{Y.G})}|G}\beta_{Z_{(\text{Y.G})}\sim G} \tag{5.35}$$

Under the standard DID assumptions, we could write $\text{ATT}_Y = \beta_{Y\sim G} - \beta_{N\sim G}$, but we want to relax these assumptions. Under an assumption of effect homogeneity, or using OLS regression as an approximation, we have that $\text{ATT}_Y = \beta_{Y\sim G|Z}$, which means that $\text{Bias}_Y = \beta_{Y\sim Z_{(\text{Y.G})}|G}\beta_{Z_{(\text{Y.G})}\sim G}$. We can also see that, $\text{DIM}_N = \beta_{N\sim G}$ and so $\text{DIM}_N - \beta_{N\sim G|Z} = \beta_{N\sim G} - \beta_{N\sim G|Z} = \beta_{N\sim Z_{(\text{N.D})}|G}\beta_{Z_{(\text{N.D})}\sim G}$. Again, there is some $m \in \mathbb{R}$ such that $\beta_{Y\sim G} - \beta_{Y\sim G|Z} = m \times (\beta_{N\sim G} - \beta_{N\sim G|Z})$. Plugging this into the expression for ATT_Y, we get Equation 5.36.

$$\text{ATT}_Y = \beta_{Y\sim G|Z} = \beta_{Y\sim G} - \beta_{Y\sim Z|G}\beta_{Z\sim G} = \beta_{Y\sim G} - m \times (\beta_{N\sim G} - \beta_{N\sim G|Z}) \tag{5.36}$$

$m = \frac{\beta_{Y \sim Z_{(\mathrm{Y.D})}|G} \beta_{Z_{(\mathrm{Y.D})} \sim G}}{\beta_{N \sim Z_{(\mathrm{N.D})}|G} \beta_{Z_{(\mathrm{N.D})} \sim G}}$ is the ratio of biases for the regression coefficients. Under the assumption of effect homogeneity for the effect of D on N or using OLS as an approximation, we see that $\beta_{N \sim G|Z} = \mathrm{ATT}_N$, $m = \frac{\mathrm{Bias}_Y}{\mathrm{Bias}_N}$, and $\beta_{N \sim G} = \mathrm{DIM}_N$. Thus, Equations 5.33 and 5.36 are equivalent. Note that we don't need to make the effect homogeneity for the effect of D on N for Equation 5.36 to be useful, however. We can simply parameterize the expression for ATT_Y with m and $\beta_{N \sim G|Z}$. Additionally, if we do not want to make such an assumption for the effect of D on Y, we can make $\beta_{Y \sim G|Z}$ our inferential target and still make use of Equation 5.36. This framing connects us back to the framework laid out in Section 5.1.

We can also define k as in Section 5.1. If the scale of N and Y differ, it will be easier to reason about the value of k than to reason about the value of m. So we now can see how we can relax both the assumption that N is a pre-treatment outcome with no treatment effect from G and the assumption that we have perfect equiconfounding. We also have seen how the latter relaxation could be parameterized in terms of $\beta_{N \sim G|Z}$ and m or k. This all can be done while maintaining an identification strategy with roots that connect easily to the standard difference in differences assumptions and identification strategy.

Moreover, we can see standard difference in differences as the special case where we assume that $\beta_{Y \sim G|Z} = 0$ and $m = 1$ or equivalently $k = \frac{\mathrm{SD}(N^{\perp G})}{\mathrm{SD}(Y^{\perp G})}$. This value of k arising could be the result of simply a fortunate balancing coincidence for scale differences and levels of confounding. But it is more likely to be interpreted as the scale of N and Y being the same (i.e., $\mathrm{SD}(N^{\perp G}) = \mathrm{SD}(Y^{\perp G})$) in addition to equiconfounding. In this way, the assumption that $k = 1$ can be thought of as less demanding than the assumption that $m = 1$. So we could consider a "new" form of difference in difference where we assume $k = 1$ but do not assume that $m = 1$. However, the above framework need not be limited to any specific choice for k. Rather, we can consider a range of plausible values for k, perhaps including the value that would bring the true effect to zero under the assumption that $\beta_{N \sim G|Z} = 0$, $k = \frac{\mathrm{DIM}_Y}{\mathrm{DIM}_N} \times \frac{\mathrm{SD}(N^{\perp G})}{\mathrm{SD}(Y^{\perp G})}$, as well as $k = 1$ and $k = \frac{\mathrm{SD}(N^{\perp G})}{\mathrm{SD}(Y^{\perp G})}$.

Transforming unequal confounding assumptions into unequal trend assumptions Finally, we show that any assumption for m corresponds to an assumption on the trend in control potential outcomes. We will show that any assumption for m can be turned into an assumption about how the trends in control potential outcomes compare. An assumption about a difference in confounding would be something like Equation 5.37. An assumption about a difference in trend would be something like Equation 5.38.

$$\begin{aligned}
&\underbrace{\mathbb{E}[Y_0|G=1]-\mathbb{E}[Y_0|G=0]}_{\text{Bias}_Y} = m \times (\underbrace{\underbrace{\mathbb{E}[N|G=1]-\mathbb{E}[N|G=0]}_{\text{DIM}_N} - \beta_{N\sim G|Z}}_{\text{Bias}_N}) \\
&\qquad \iff \mathbb{E}[Y_0|G=1] = \mathbb{E}[Y_0|G=0] + m \times (\mathbb{E}[N|G=1]-\mathbb{E}[N|G=0]-\beta_{N\sim G|Z})
\end{aligned} \tag{5.37}$$

$$\begin{aligned}
&\underbrace{\mathbb{E}[Y_0|G=1]-\mathbb{E}[N|G=1]}_{\text{Trend}_{G=1}} = w \times (\underbrace{\mathbb{E}[Y_0|G=0]-\mathbb{E}[N|G=0]}_{\text{Trend}_{G=0}}) \\
&\qquad \iff \mathbb{E}[Y_0|G=1] = \mathbb{E}[N|G=1] + w \times (\mathbb{E}[Y_0|G=0]-\mathbb{E}[N|G=0])
\end{aligned} \tag{5.38}$$

For this exposition, we assume $\beta_{N\sim G|Z}=0$ for simplicity. Both Equations 5.37 and 5.38 provide an expression for $\mathbb{E}[Y_0|G=1]$, which, in the DID setting, is unobserved while all the other components of the expressions are estimable from the data. We can set these two expressions for $\mathbb{E}[Y_0|G=1]$ equal to each other and then solve for w to get the trend assumption that corresponds to our assumption on m.

$$\begin{aligned}
&\mathbb{E}[Y_0|G=0] + m \times (\mathbb{E}[N|G=1]-\mathbb{E}[N|G=0]) \\
&= \mathbb{E}[N|G=1] + w \times (\mathbb{E}[Y_0|G=0]-\mathbb{E}[N|G=0])
\end{aligned} \tag{5.39}$$

$$\begin{aligned}\iff w &= \frac{\mathbb{E}[Y_0|G=0] + m \times (\mathbb{E}[N|G=1] - \mathbb{E}[N|G=0]) - \mathbb{E}[N|G=1]}{\mathbb{E}[Y_0|G=0] - \mathbb{E}[N|G=0]} \\ \iff m &= \frac{\mathbb{E}[N|G=1] + w \times (\mathbb{E}[Y_0|G=0] - \mathbb{E}[N|G=0]) - \mathbb{E}[Y_0|G=0]}{\mathbb{E}[N|G=1] - \mathbb{E}[N|G=0]}\end{aligned} \tag{5.40}$$

We could similarly write m in terms of w, if we preferred to make an assumption on w, the difference in the trend, and wanted to know what this implies about the unequal confounding. Additionally, we know that $m = k \times \frac{\text{SD}(Y^{\perp G})}{\text{SD}(N^{\perp G})}$ so that we can make an assumption on k and see what this implies about w, the trend, or make a trend assumption on w and see what this implies about k.

5.4 Examples

5.4.1 National Supported Work Demonstration

Let us consider a substantive example. The National Supported Work Demonstration (NSW) was a job training program aimed at helping disadvantaged worker build basic skills in the 1970s. The program randomly assigned participants to training positions. These consisted of a group that received the benefits of the program (the treatment group) and a group that did not receive any benefits (the control group). (LaLonde, 1986) The data also include a non-experimental dataset in which individuals not from the NSW program but from the Panel Study of Income Dynamics (PSID) are added as additional control units. The data that we work with is "a subset that consists of males with three (2 pre-training and 1 post-training) years of earnings data (the outcome of interest). This subset has been considered in (Dehejia and Wahba 1999; Smith and Todd 2005; Firpo 2007)." (Callaway, 2019) The data can be found in the "`qte`" `R` package by running the command `data(lalonde,package = "qte")`. We use the non-experimental dataset (called `lalonde.psid`) to illustrate the methods discussed in this paper and then compare these results with the estimates from the experimental data (called `lalonde.exp`).[21] The difference in means estimate of the treatment effect from the experimental data is 1,795.55 higher 1978 earnings in dollars for the treated group relative to the control group. The estimate is 1,693.12 when we control for observed covariates.

We use 1974 earnings (`re74`) as a placebo outcome for 1978 earnings (`re78`). The treatment (`treat`) is experiencing the job training program. We also use the set of covariates (X) in the data including age, education (years and degree or not), marital status, and demographic indicators. We recognize that there are unobserved covariates (Z) that we wish we could condition on but cannot. Ideally, we would like to estimate the regression `lm(re78 ~ treat + X + Z)`. Due to the unobserved confounders, we will not be able to point identify the coefficient on `treat` from this "long" regression. Instead, we partially identify $\beta_{Y\sim D|X,Z}$ with the approach proposed by the present paper and leverage information on 1974 earnings as a placebo outcome. We note that it is likely that 1974 earnings influence 1978 earnings. However, Chabé-Ferret (2017); Ham and Miratrix (2022) warn against conditioning on pre-treatment outcomes in a difference-in-differences type approach. We follow that guidance here for simplicity. Thus we use the approach from Section 5.1.2. Since the placebo outcome is pre-treatment earnings, it is reasonable to assume that the treatment has no causal effect on the the placebo outcome. Therefore, we consider that $\beta_{P\sim D|Z,X} = 0$. We might think that a straightforward difference in differences approach is possible. However, we might worry that there is unequal levels of confounding between 1974 earnings and 1978 earnings. For example, minimum wages increased 32% from 1974 to 1978. (Department of Labor, 2023) The treatment group (comprised of disproportionately lower earners) likely experienced a larger increase in earnings from the rise in minimum wage than did the control group. Thus, the difference in baseline earnings between the treated group and control group in 1978 likely smaller than in 1974. So we might consider $k < 1$. On the other hand, pre and post-treatment earnings are measures of the same variable, so their levels of confounding

[21] Since this data includes a randomized experiment, we can also estimate the effect using the experimental data to get an unbiased estimate of the true treatment effect. This experimental estimate can be used to back out the true value for k, assuming that $\beta_{N\sim D|Z,X} = 0$.

should be somewhat similar. Suppose that $k > 0.5$. We also might worry that high inflation in the 1970s could lead to scale differences between 1974 earnings and 1978 earnings (our placebo and actual outcomes).

We carryout the partial identification analysis using the "`placeboLM`" `R` package the authors built for this purpose. `placeboLM` is currently under development; a software paper will accompany its release.

```
# install.packages("devtools")
devtools::install_github("Adam-Rohde/placeboLM")
library(placeboLM)
```

This package can be used in a manner reminiscent of the `lm()` function in `R`. Arguments to the `placeboLM` function include the dataset, the outcome variable, the treatment variable, the placebo variable (and whether we want to think of it as a placebo outcome or treatment), the direction or presence of the edge from the treatment (D) to the placebo (P), the direction or presence of the edge from the placebo (P) to the outcome (Y), the observed covariates, and finally the ranges of the partial identification parameters that we want to consider. These ranges can be set to be the range that is plausible or something more conservative than that. The first partial identification parameter is k, which captures the ratio of the level of confounding of D and Y, conditional on P, to the level of confounding of D and P, after accounting for scale differences. We can think of this as the D, Y confounding as a percentage of the D, P confounding, where scale differences between P and Y have been accounted for. The second partial identification parameter is $\beta_{P \sim D|Z,X}$, which captures the direct causal relationship between D and P and we already noted should be zero. This parameter is referred to as `coef_P_D_given_XZ` in the software. For this example, we set the partial identification parameter ranges to be conservative and assume that the placebo (1974 earnings) directly causes the outcome (1978 earnings).

```
plm = placeboLM(
  data = "lalonde.psid",
  outcome = "re78",
  treatment = "treat",
  placebo_outcome = "re74",
  observed_covariates = c("age", "education", "black", "hispanic",
                          "married", "nodegree"),
  partialIDparam_minmax = list(
                        k = c(-2,2),
                        coef_P_D_given_XZ = c(-15000,15000)) )
```

Running the `placeboLM` function provides us with a quick summary of the setting that we have chosen and the regressions that we will estimate as part of the partial identification framework. We can then use the three main analysis functions of the `placeboLM` package to do partial identification for $\beta_{Y \sim D|X,Z}$, which is the regression coefficient that we would get if we could include the unobserved confounders in our regression. The first is `placeboLM_table()`. We can specify which percentiles of the range of plausible values for the partial identification parameters we would like to report estimates for. The rows with these estimates are labelled "`Grid`". We get bootstrapped standard errors and confidence intervals in addition to the estimates. Finally, we also get the estimates under common point identification assumptions: selection on observables (SOO), standard difference in differences (DID), and DID assuming $k = 1$.

```
set.seed(0)
placeboLM_table(plm, n_boot = 1000, ptiles = c(0.25,0.5,0.75), alpha = 0.05)
```

The second key partial identification function is `placeboLM_contour_plot()` which provides a contour plot of the estimates as we vary the two partial identification parameters

Table 5.1: Partial identification table for National Supported Work example with 1974 earnings as placebo outcome.

	k	coef_P_D_given_XZ	Estimate	Std. Error	CI Low	CI High
SOO	0	0	-5928.11	822.57	-7563.28	-4337.09
DID	0.845	0	1718.01	852.98	27.35	3393.96
k=1 DID	1	0	3115.31	882.42	1337.4	4783.3
Grid	-1	-7500	-6100.92	1375.32	-8843.01	-3539.27
Grid	-1	0	-14971.53	1379.99	-17615.87	-12160.46
Grid	-1	7500	-23842.13	1484.07	-26566.69	-20663
Grid	0	-7500	-5928.11	809.4	-7473.88	-4304.65
Grid	0	0	-5928.11	842.29	-7549.07	-4215.56
Grid	0	7500	-5928.11	832.39	-7523.88	-4262.12
Grid	1	-7500	-5755.3	871.06	-7441.71	-4051.86
Grid	1	0	3115.31	926.64	1389.42	4907.9
Grid	1	7500	11985.91	984.87	10176.39	13963.14

within the ranges we originally provided. The plot also includes the SOO and DID estimates for reference. We highlight the contour on which estimates are zero. See Figure 5.8.

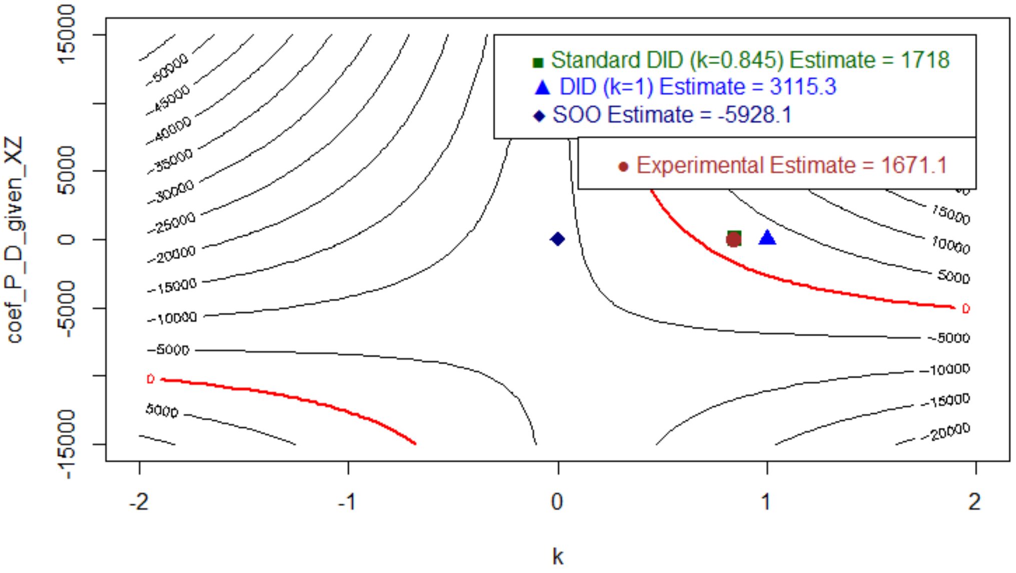
```
placeboLM_contour_plot(plm, gran = 100)
```

The final key partial identification function is `placeboLM_contour_plot()` which provides a line plot of the estimates as we vary one of the two partial identification parameters. We can specify percentiles of the other partial identification parameter and have a line plot generated for each. Since we believe `coef_P_D_given_XZ` is zero in this example, we focus only on this case. The line plot also includes 95% confidence intervals. See Figure 5.7.

```
placeboLM_line_plot(plm, bootstrap=TRUE, n_boot=1000, ptiles = c(0.5),
                    focus_param = "k", ptile_param = "coef_P_D_given_XZ",
                    gran= 10, alpha = 0.05)
```

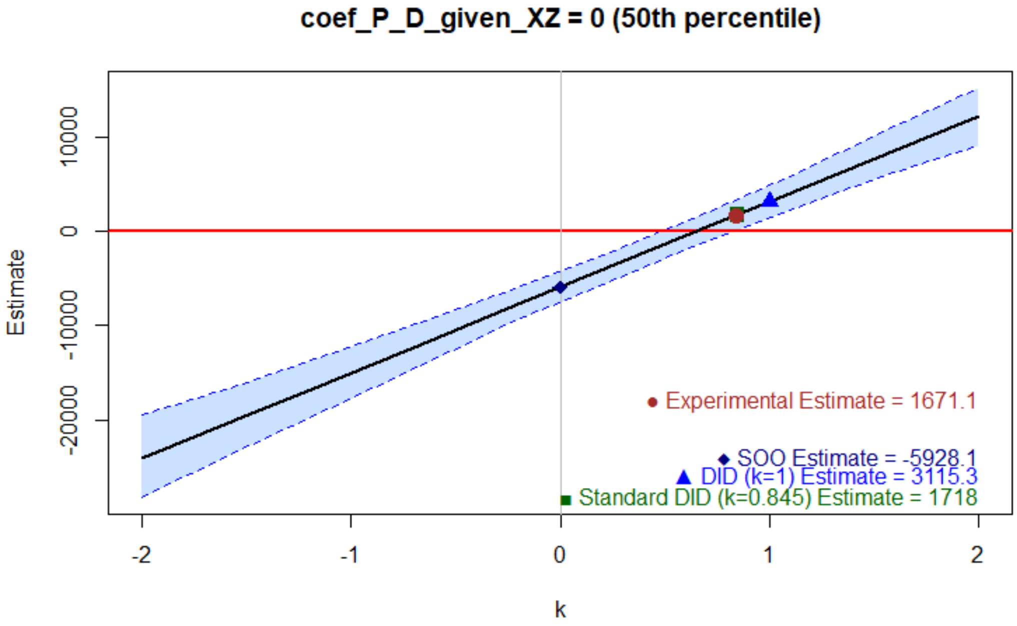
The contour plot shows how the effect estimate for $\beta_{Y\sim D|X,Z}$ changes as k and $\beta_{P\sim D|Z,X}$ change. We've noted that it is likely that $\beta_{P\sim D|Z,X} = 0$. The line plot shows us how the

Figure 5.6: Contour plot for National Supported Work example with 1974 earnings as placebo outcome.



effect estimate for $\beta_{Y\sim D|X,Z}$ changes as k changes, under the assumption that $\beta_{P\sim D|Z,X} = 0$. The SOO estimate is negative. This is not the sign we would intuitively expect, since training should probably boost earnings. The standard DID estimate, however, is positive and implicitly makes an assumption that the bias for the treatment and 1978 earnings is 85% of the bias for the treatment and 1974 earnings, after accounting for scale differences. The estimate under this assumption is an increase of 1,718.01 dollars as a result of experiencing the job training. The $k = 1$ DID estimate is also positive and provides an estimate of an increase of 3,115.31 dollars as a result of experiencing the job training. $k = 1$ corresponds to an assumption that the bias for the treatment and 1978 earnings is 100% of the bias for the treatment and 1974 earnings. Under the assumption that $\beta_{P\sim D|Z,X} = 0$, the experimental estimate (1,693.12) implies that the "true" value for k is about 0.84 or that the bias for the treatment and 1978 earnings is 84% of the bias for the treatment and 1974 earnings, after accounting for scale differences.

Figure 5.7: Line plot for National Supported Work example with 1974 earnings as placebo outcome.



Without knowledge of the experimental estimate, however, we might consider the line plot and make an argument that the bias for the treatment and 1978 earnings and the bias for the treatment and 1974 earnings have the same sign. This would put us on the right half of the line plot. From our previous discussion, we might consider the range $k \in [0.5, 1]$ as plausible. This translates to non-negative effect estimates - either our effect is not statistically significant or it is positive and statistically significant. This is proposed for illustration only. Subject matter experts could debate this range.

5.4.2 Zika Virus

In our second application, we investigate the effect Zika virus had on birth rates in Brazil. Zika virus can cause birth complications. Brazil was heavily impacted by the outbreak of Zika virus in 2015. We use data that has also been analyzed by Taddeo et al. (2022); Tchetgen et al. (2023a). The virus had a dramatic impact on the state of Pernambuco but zero cases were reported in the state of Rio Grande do Sul. As in previous works, we treat the former state as the treated group and the later as the control group. We use municipality level overall birth rates per 1000 people in 2016 as our primary outcome of interest. We use municipality level overall birth rates per 1000 people in 2014 as our placebo outcome. We refer curious readers to the previously cited paper for additional discussion of the virus and its spread in Brazil. We make the reasonable assumption that the treatment has no direct causal relationship with the pre-treatment outcome. We use the data from Amorim (2022) and run the following application of `placeboLM`. We can get a full picture of partial identification just using `placeboLM_line_plot()`.

```
plm = placeboLM(
  data = "data_reshape",
  outcome = "Rate_2016",
  treatment = "trt",
  placebo_outcome = "Rate_2014",
  partialIDparam_minmax = list(k = c(0,2), coef_P_D_given_XZ = c(-10,10)) )
```

```
set.seed(0)
placeboLM_line_plot(plm, bootstrap=TRUE, n_boot=1000, ptiles = c(0.5),
                    focus_param = "k", ptile_param = "coef_P_D_given_XZ",
                    gran= 10, alpha = 0.05)
```

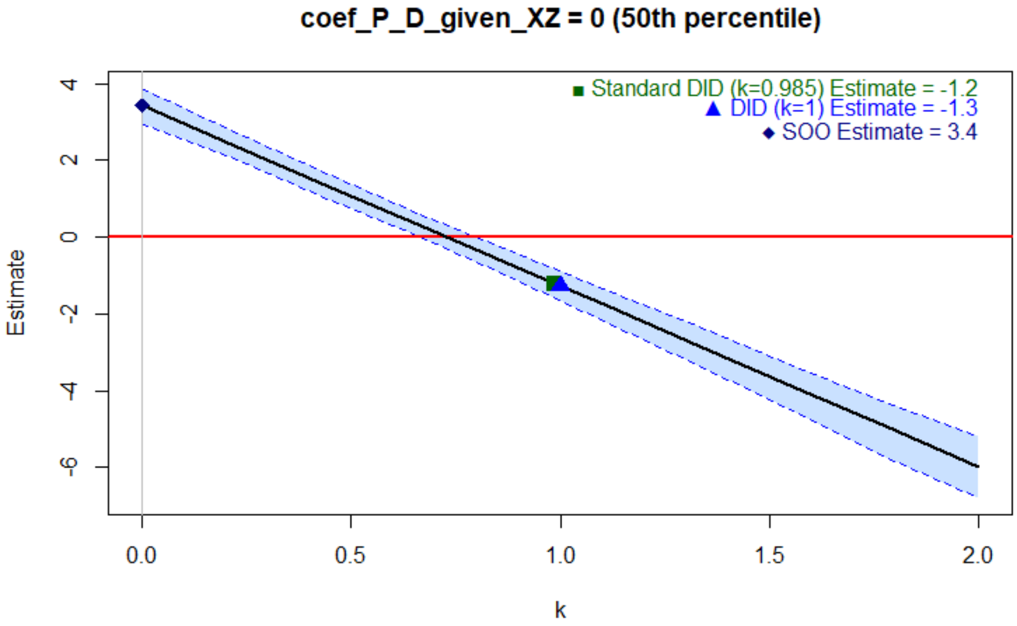
We find that the naive difference in means estimate is that Zika is associated with an rise in the birth rate by 3.4 births per 1000 people. However, the two difference in difference estimates are negative and indicate that there were about 1.2 to 1.3 fewer births per 1000 people in Pernambuco as a result of the Zika virus. We can also see that if the level of confounding of 2016 birth rate and treatment is above approximately 70% of the level of confounding of 2014 birth rate and treatment, then the effect estimates are always negative. If the relative level of confounding is something like 120%, the effect estimate may be around 2. It is reasonable to believe that the relative level of confounding is not too far from 100%. So we might conclude that the Zika virus reduces birth rate or at least does not increase it. These findings are consistent with the results in Taddeo et al. (2022); Tchetgen et al. (2023a); but using `placeboLM` and the partial identification framework outlined in the present paper builds in a skepticism about whether untestable assumptions like those made in Tchetgen et al. (2023a) actually hold. Rather than hold to a single assumption, we can explore the range of plausible possibilities. Additionally, we could conceive of the state indicator, which is used as a proxy for whether or not the population was exposed to Zika virus, as a direct cause of 2014 birth rates, in that the two states likely differ in important ways that lead to divergent birth rates in 2014. Such a view could easily be accommodated in our framework by considering values for `coef_P_D_given_XZ` other than zero. This would, on the other hand, be a violation of key assumptions in Tchetgen et al. (2023a).

5.5 Discussion

5.5.1 Related work

For a recent example of a placebo approach employing equiconfounding, see Ye et al. (2022). Lipsitch et al. (2010) provide an early discussion of the potential of using placebo treatments and outcomes. Arnold et al. (2016); Arnold and Ercumen (2016) discuss the potential to use placebos to address sample selection and measurement bias as well as bias in randomized experiments. Shi et al. (2020) review many established procedures and recent developments in this growing literature. An important recent area of research has been dubbed "proximal" causal inference. Tchetgen et al. (2020) provides an introduction. This line of work leverages "double-negative control designs" (Shi et al., 2020) in which both a placebo outcome and a placebo treatment are present. Under certain assumptions and conditions, papers like Miao

Figure 5.8: Line plot for Zika virus example with 2014 birth rate as placebo outcome.



et al. (2018, 2020); Mastouri et al. (2021) show how causal effects are non-parametrically identifiable using both of these placebos. While this work is very promising, it requires two placebos, certain null effect assumptions (i.e., "perfect" placebos), and regularity conditions.[22] Tchetgen et al. (2023a) provides a related approach for the case when only a single placebo is available. In this approach, the placebo variable must be associated with the control potential outcome and must be associated with the treatment only through the control potential outcome; there cannot be any causal relationship between the treatment and the placebo other than those that run through the control potential outcome for the primary outcome. This approach involves untestable assumptions. It also requires that "for any variation in the [primary outcome], there is corresponding variation in the [placebo]." This may preclude us from leveraging placebos in settings where they could be informative. For example, the

[22] In Section 5.1.4, we show a similar point identification result for regression when both a "perfect" placebo treatment and a "perfect" placebo outcome are available. We also relax the "double placebo" setting to allow for imperfect placebos.

assumptions are violated if the placebo has a common cause with the treatment that is not also a cause of the primary outcome or if the placebo plays a direct role in the treatment assignment mechanism. Additionally, if we have a placebo treatment, this approach may not apply. Our proposed approach, however, is applicable in these settings.

Additionally, Sofer et al. (2016) point out that placebo outcome approaches can be seen as generalizations of the familiar difference in differences (DID) framework. DID uses a pre-treatment verison of the outcome as a placebo outcome and relies heavily on a "parallel trends" assumption that the baseline trend in the outcome is the same over time for the group that gets treated and the group that does not. (Angrist and Pischke, 2008) Parallel trends is, in fact, an assumption of no unobserved time-varying confounding; meaning that DID approaches have not circumvented the problem of unobserved confounding. Sofer et al. (2016) also show how the parallel trends assumption is equivalent to an equiconfounding assumption. Recently, there have been several attempts to relax the parallel trends assumption within the context of DID or to improve on the current practice of testing for parallel trends in the pre-treatment period. For example, see Manski and Pepper (2018); Bilinski and Hatfield (2018); Freyaldenhoven et al. (2019); Keele et al. (2019); Ryan et al. (2019); Ye et al. (2020); Gibson and Zimmerman (2021); Rambachan and Roth (2021). While these are all useful contributions, they may not be as easy to use as applied researchers might hope and make assumptions that researchers might be interested in relaxing. Gibson and Zimmerman (2021) points out that violations of parallel trends can be considered using an application of the omitted variable bias framework from Cinelli and Hazlett (2020) in a first difference regression. However, this approach is limited to the DID setting in which the pre-treatment outcome is directly comparable to the post-treatment outcome. This approach (and others like Rambachan and Roth (2021)) also do not allow for a relaxation of the assumption that the treatment has no direct causal relationship with the pre-treatment outcome. Additionally, this approach requires that researchers observe the pre-treatment outcome and the post-treatment outcome for the same set of units; that is, it requires panel data. Tchetgen et al. (2023b) discusses an alternative to the parallel trends assumption which is a particular form of equiconfounding. Finally, the DID approach in general requires data on a pre-treatment version of the outcome, which may not always be available, making the placebo paradigm an attractive generalization.

5.5.2 Conclusion

We have explored how the analysis of "omitted variable bias" in regression provides a simple yet powerful way to leverage information from placebo treatments and outcomes for partial identification of causal effects or approximations of them. We have seen that the framework we developed allows for violations of assumptions made in traditional difference in differences, as well as in standard negative control outcome (and exposure) settings. These can be viewed as special cases of our approach. Additionally, our framework requires only a single placebo to be observed and does not require assumptions of null effects, certain conditional independencies, or unique solutions to integral equations as in Tchetgen et al. (2020, 2023a). Moreover, our approach is applicable to various causal graphs. We obtain these gains by relying on linear regression machinery but do not require any distributional assumptions. Relaxations of this are explored in Appendix D.2. We also saw that the forthcoming `placeboLM` R package facilitates easy application of the proposed framework.

APPENDIX A

Appendix for Chapter 2

Here we provide technical details and prove the results found in the main text. First, we briefly discuss why we work with potential outcomes. We then introduce a series of definitions. These are followed by a series of lemmas. Then we state our main results in a set of theorems that follow directly from the lemmas. We conclude with discussions of IPW estimation and when conditioning on a collider creates an association.

A.1 Sample selection, the do-operator, and potential outcomes

Let us linger on the choice to use potential outcomes as opposed to the do-operator (Pearl, 1995, 2009) in our discussion. We discuss post-treatment selection, selection as a mediator, selection as the child of a mediator, and all other major roles that selection can play in the structure of causal models. In the case of non-post-treatment selection, potential outcomes will typically have the same interpretation as the do-operator. In the case of selection as a mediator and selection as a descendant of a mediator, the causal effects of interest are typically defined using potential outcomes. (Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2011; VanderWeele and Vansteelandt, 2009; Richiardi et al., 2013) So here potential outcomes notation is really the natural and traditional choice. In the case of post-treatment selection, where selection is not a mediator or a descendant of a mediator, $p(Y_d|S=1)$ is usually of interest and not $p(Y|do(D=d), S=1) = p(Y_d|S_d=1)$.[1] As Pearl (2015a) states, ”By the

[1] $p(Y|do(D=d), S=1) = \frac{p(Y, S=1|do(D=d))}{p(S=1|do(D=d))} = \frac{p(Y_d, S_d=1)}{p(S_d=1)} = p(Y_d|S_d=1)$. (Pearl, 2014) In this case, S_d is the potential selection value when the treatment variable D takes the value that we are investigating in Y_d.

counterfactual query $Q_c[= p(Y_d|S = 1)]$ we mean: Take all units which are currently at level $[S = 1]$, and ask what their Y would be had they been exposed to treatment $[D = d]$. This is different from $Q_{do} = [p(Y|do(d), S = 1)]$, which means: Expose the whole population to treatment $[D = d]$, take all units which attained level $[S = 1]$ (post exposure) and report their Y's." Further, Pearl (2015a) states "Q_{do} is rarely posed as a research question of interest, probably because it lacks immediate causal interpretation. It serves primarily as an auxiliary mathematical object in the service of other research questions. ... I have not seen Q_{do} presented as a target query on its own right." For non-post-treatment selection, $p(Y_d|S = 1) = p(Y|do(d), S = 1)$, since $S_d = S$. In our context, we will typically be interested in quantities of the type of $p(Y_d|S = 1)$, which tell us the distribution of outcomes for units that were selected in reality, had they been exposed to treatment $D = d$.

A.2 Definitions

Definition A.2.1 (SCM (adapted from Pearl (2009))). *A structural causal model, M, has the following parts*

1. *U is a set of background variables determined by exogenous factors;*
2. *V is a set $\{V_1, V_2, \dots, V_n\}$ of variables determined by variables in the model;*
3. *F is a set $\{f_1, f_2, \dots, f_n\}$ of functions that map $f_i : U_i \cup PA_i \rightarrow V_i$, where $U_i \subset U$ and $PA_i \subset V \backslash V_i$ and the entire set F forms a mapping from U to V. That is, each f_i assigns a value to V_i that depends on the values of a select set of variables in $V \cup U$ ($v_i = f_i(pa_i, u_i)$), and the entire set F has a unique solution $F(u)$.*
4. *$p(u) = \prod p(u_j)$ is a probability function defined over the domain of U.*

Definition A.2.2 (Sub-Model (adapted from Pearl (2009))). *Let M be a causal model, D be a set of variables in V, and d a particular realization of D. A submodel M_d of M is the causal model M_d, where F is replaced with F_d, which is formed by deleting the functions for the variables in D and replacing them with constant functions $D = d$.*

Definition A.2.3 (Potential Outcome (adapted from Pearl (2009))). *Let D and Y be two subsets of variables in V. The counterfactual values of Y when D had been set to d, written Y_d, is the solution for Y of the set of equations F_d, given the realized values of the background variables, U.*

Definition A.2.4 (Causal Graph (adapted from Shpitser et al. (2010), also see Pearl (1988, 2009))). *A SCM induces a causal graph in the following way. Each variable in the model is represented by a node. A node corresponding to variable V_i has edges pointing to it from every variable whose value is used to determine the value of V_i by the function f_i. Exogenous variables have no edges pointing to them. A causal graph is an I-map (see Definition A.2.11 below) for $p(v)$.*

Definition A.2.5 (Path). *A path is a sequence of edges in G where each pair of adjacent edges in the sequence share a node, and each such shared node can occur only once in the path.*

Definition A.2.6 (Causal Path). *A causal path from D to Y is a path from D to Y on which all edges are directed and point away from D and toward Y.*

Definition A.2.7 (Proper Causal Path (Shpitser et al., 2010))). *Let D, Y be sets of nodes. A causal path from a node in D to a node in Y is called proper if it does not intersect D except at the end point.*

Definition A.2.8 (Non-Causal Path). *A non-causal path is a path that is not a causal path.*

Definition A.2.9 (Parents, Ancestors, and Descendants). *Parents of node X are the nodes in the graph from which an edge points directly to X. An ancestor of X is any node which has a causal path to X. A descendant of X is any node which X has a causal path to.*[2]

Definition A.2.10 (d-Separation and Blocking (adapted from Pearl (2009))). *Two sets of nodes, D, Y, in a graph G are said to be d-separated by a third set, Z, if every path from any*

[2] We do not consider a node to be a descendant of itself.

node $D_0 \in D$ to any node in $Y_0 \in Y$ is blocked. A path is blocked by Z if either [1] some W is a collider on the path between D, Y and $W \notin Z$ and the descendants of W are not in Z or [2] W is not a collider on the path but $W \in Z$.

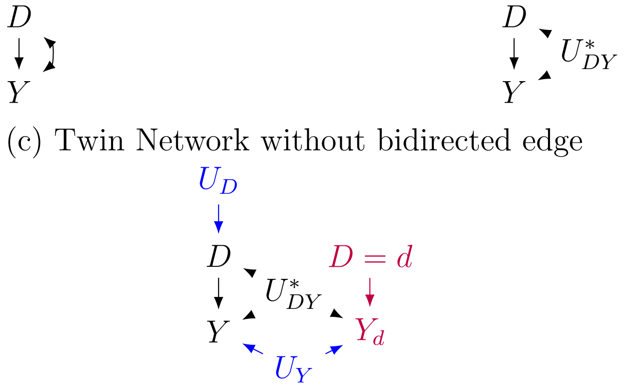
Definition A.2.11 (I-map (adapted from Pearl (1988))). *A causal graph G is said to be an I-map of a dependency model M if every d-separation condition displayed in G corresponds to a valid conditional independence relationship in M. That is, for every set of three nodes X, Y, and Z, if Z d-separates X from Y in G, then X is independent of Y given Z.*

Shpitser et al. (2010) discuss a graphical representation called latent projections of causal graphs that contain both directed and bidirected edges. Latent projections allow us to exclude latent variables in convenient ways. Specifically, they include a node for every observed variable. However, two observable nodes A and B are connected by a directed edge only when any and all intervening variables between A and B are latent. Also, A and B are connected by a bidirected edge when there is a path from A to B that is not d-separated that starts with an edge pointing into A and ends with an edge pointing into B and all the nodes on this path are latent other than the end points. As Shpitser et al. (2010) point out, latent projections retain all d-separation statements from the original graph. We will also allow for such latent projections to be used to simplify graphs. For our purposes, we do not allow sample selection to be treated as a latent variable and so it should always be included as a separate node in the graph.

Definition A.2.12 (Twin-Network (adapted from Shpitser et al. (2010))). *The twin network graph, N, (Balke and Pearl, 1994b,a) displays counterfactual independence among two possible worlds, the pre-intervention world which is represented by the original graph G, and the post-intervention world, which is represented by the graph $G_{\overline{D}}$ (a copy of G with the edges pointing into D deleted and D replaced with $D = d$). The twin network is an I-map for the joint counterfactual distribution $p(v, v_d)$, where V is the set of all observables, and V_d is the set of all observable variables after the intervention $do(D = d)$ was preformed. The observable nodes* *in these two graphs share the U variables, to signify a common history of these worlds up to the point of divergence due to $do(D = d)$. We add the additional refinement from Shpitser and Pearl (2007) where node copies of all non-descendants of D in G and $G_{\overline{D}}$ are merged in the twin network graph (since such nodes are the same random variable in both the pre and post intervention worlds).*

In our proofs, we will consider causal graphs and twin networks in which each bidirected edge between nodes A and B is replaced with a node U^*_{AB} that is a common cause of the two nodes that were connected with the bidirected edge and points to each of A and B. This replacement does not change d-separations from the original graph. See Figure A.1 for a simple example. Correa et al. (2018) make a similar alteration to the causal graphs they consider.

Figure A.1: Twin network with no bidirected edges



(a) DAG with bidirected edge

(b) DAG without bidirected edge

(c) Twin Network without bidirected edge

Definition A.2.13 (Colliders). *A collider is a node in a causal graph into which two (or more) arrow heads point. For nodes A, B, C, let C be a collider between A and B if it appears in the following sub-path of the causal graph: $A \to C \leftarrow B$.*

Definition 1 (Internal Selection Graph, G^+_S). Let G be the DAG induced by a SCM.

1. Create G_S by adding an appropriately connected binary selection node, S.
2. Draw a circle around S to clearly indicate that we must limit our analysis to $S = 1$.
3. Add to G_S any node which is a parent of the treatment or a parent of a descendant of the treatment. (U_S, the background factors contributing to selection, can be excluded.)
4. Add a dashed undirected edge between all variables between which S is a collider or an ancestor of S is a collider. We will call these dashed, undirected edges *bridges*.

Call the resulting graph an *internal selection graph, G_S^+*.

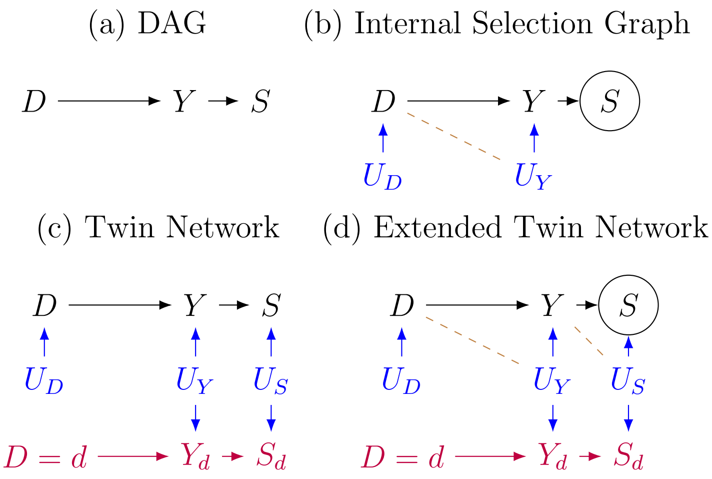
(This definition is similar to the "modified extended diagram" in Daniel et al. (2012).)

Definition A.2.14 (Extended Twin-Network). *An extended twin network, N_S^+, is a twin network, N_S, containing an appropriately connected pre-intervention binary selection node, S, and any corresponding post-intervention versions of it, where we add bridges between all variables between which the pre-intervention S is a collider or an ancestor of pre-intervention S is a collider. (Note that pre and post-intervention versions of S are assumed to have been added to both N_S and N_S^+; we don't use a subscript to indicate this here.) It is easy to see that, like a twin network, an extended twin network displays counterfactual independence among two possible worlds, the pre-intervention world which is represented by the original graph G_S^+, and the post-intervention world, which is represented by the graph $(G_S)_{\overline{D}}$.*

Extended twin networks are useful for the same reason that internal selection graphs are useful. There can be purely statistical relationships between variables in the sample that are not captured in regular twin networks. See Figure A.2. As we saw in the main text, bridges do not create colliders, since they are graphical representations of conditioning on sample selection when it is a collider. So bridges do not alter the underlying fully directed graph. Since the addition of bridges does not create any colliders, d-separation and blocking retain their definition in internal selection graphs and extended twin networks. See Lemmas 4 and 5 that show how d-separation (using the same definition) in internal selection graphs and extended twin networks corresponds to d-separation in causal graphs and twin networks.

As a result, we can then get independence statements by reasoning about internal selection graphs and extended twin networks.

Figure A.2: Example of Extended Twin Network



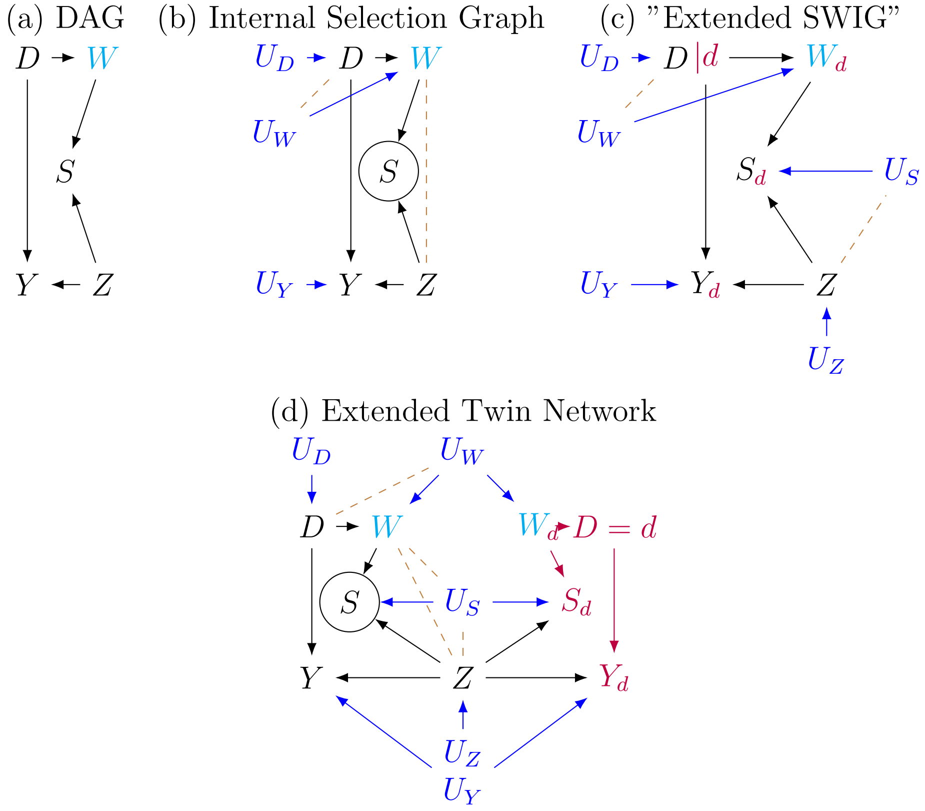
Twin network graphs can become pretty complicated, even when the original causal graph only contains three nodes.[3] This is what makes graphical criteria like the one presented in this paper attractive for simplifying the analysis that leads to ignorability statements. The internal selection graph maintains only the necessary elements of the extended twin network that allow us to use the internal selection criterion to see when conditional ignorability is possible. We are not advocating that researchers actually work with extended twin networks

[3] Richardson and Robins (2013a,b) also introduce a graphical approach to visualizing how post-intervention world variables relate to pre-intervention world variables called Single World Intervention Graphs (SWIGs). SWIGs are often simpler than twin networks. However, they do not provide exactly the same picture of both the pre-intervention world and the post-intervention world that twin networks do; for example, pre-intervention world post-treatment variables do not appear in SWIGs, though they can be used to block generalized non-causal paths between the treatment and potential outcome of interest. See Figure A.3. Here, W can be conditioned on to block the open generalized non-causal path between D and Y_d but the pre-intervention world W does not appear in the SWIG, while it does appear in the twin network. The pre-intervention selection node is also missing from the SWIG. Finally, the path $D \rightarrow W \cdots Z \rightarrow Y_d$ is also missing from the SWIG; we do not want to have W_d touch any bridges since it is not actually a parent of the pre-intervention selection node. As such, we use twin networks for our discussion. This example also demonstrates why the criterion in Daniel et al. (2012) is less general than we might want, since those authors do not allow any adjustment for post-treatment variables. We believe that internal selection graphs and ISAC provide the simplest approach for practitioners to analyze sample selection and internal validity, by only slightly extending the causal graphs that they are used to seeing.

themselves. We discuss extended twin networks in our proofs only. We advocate using internal selection graphs, which are usually much simpler than twin networks and extended twin networks, and the internal selection adjustment criterion for determining ignorability.

Figure A.3: Twin Networks versus Single World Intervention Graphs (SWIGs)

In this example, W can be used to block generalized non-causal paths between D and Y_d; however, only the counterfactual world W_d appears in the "extended SWIG." The actual world W and the counterfactual world W_d both appear in the twin network.



(a) DAG

(b) Internal Selection Graph

(c) "Extended SWIG"

(d) Extended Twin Network

Definition A.2.15 (Paths and Generalized Non-Causal Paths)**.** *We revise Definition A.2.5 to state that a path is a sequence of edges in G_S^+ or N_S^+ where each pair of adjacent edges in the sequence share a node, and each such shared node can occur only once in the path, where we allow the edges to be bridges, as well as directed edges. A generalized non-causal path is a path that is not a causal path.*

Definition A.2.16 (Route (adapted from Shpitser et al. (2010))). *A route from D to Y in a graph, G_S^+ or N_S^+, is a sequence of edges, where each pair of adjacent edges share a node, the unshared node of the first edge is D, and the unshared node of the last edge is Y. (Shared nodes can occur more than once.) A route is d-separated if the same triples are blocked as in the definition of d-separation above. The difference between a route and a path is that paths cannot contain duplicate nodes while routes can. Note that we allow edges to be bridges.*

Definition A.2.17 (Direct Route (adapted from Shpitser et al. (2010))). *Let π be a route from D to Y in G_S^+ or N_S^+. Label each node occurrence in the route π by the number of times the node has already occurred earlier in π. A direct route π^* is a sub-sequence obtained from π inductively as follows:*

- *The first node in π^* is the first node in π with the largest occurrence number.*
- *If the kth shared node in π^* (and the mth node in π) is (X_i, r), and $X_i \neq Y$, let the $k+1$th node in π^* be (X_j, n), where X_j is the $m+1$th node in π, and n is the largest occurrence number of X_j in π.*

Definition 2 (Internal Selection Adjustment Criterion (ISAC)). A set of nodes Z in G_S^+ satisfies the internal selection adjustment criterion relative to D (treatment) and Y (outcome) if

1. No element of Z lies on or is a descendant of a node that lies on a causal path originating from D and arriving at Y. Note that an element of Z could be a descendant of D itself, if it is not on a causal path from D to Y. Note also that elements of Z should not be on, or descendants of nodes on, causal paths even if S is also on the causal path.
2. Z blocks every generalized non-causal path between D and Y that does not pass through S. Note that generalized non-causal paths passing through M, when S is a descendant of mediator M, on which M is an ancestor of Y also do not need to be blocked, assuming the previous condition is not violated. Further, M should not be a member of Z from the previous condition.

Definition 3 (Generalization Criterion (GC))**.** A set of nodes Z in G_S^+ satisfies the generalization criterion relative to D (treatment) and Y (outcome) if

- Z satisfies ISAC relative to D and Y and
- $Z_{\text{Ext}} \subset Z$ blocks all causal and generalized non-causal paths between Y and S in G_S^+ other than those that end in a causal path from D to Y.

This definition is a translation of Definition 8 from Correa et al. (2018).

A.3 Lemmas

Lemma 1 (adapted from Shpitser et al. (2010); Pearl (1988))**.** *Let G be a causal graph. Then any model M with a distribution $P(u, v)$ inducing G, if A is d-separated from B by C in G, then A is independent of B given C, which we write $A \perp\!\!\!\perp B|C$ in $P(u, v)$.*

Lemma 2 (adapted from Shpitser et al. (2010))**.** *For every route π in G_S^+, the direct route π^* is a path. Moreover, if π is unblocked, then π^* is unblocked.*

A.3.1 Selection unrelated to mediators

Lemma 3. *If Z satisfies ISAC in G_S^+ relative to D and Y and S is not a mediator or descendant of a mediator between D and Y, then Z d-separates D and Y_d in N_S^+.*

Proof. We very closely follow the structure of the proof of Theorem 4 of Shpitser et al. (2010). We will show the contrapositive: assuming that we are conditioning on Z, an unblocked path from D to Y_d in N_S^+ implies that ISAC is violated in G_S^+ relative to D and Y. We will proceed in the following manner:

1. Discuss the structure of π, an unblocked path from D to Y_d in N_S^+.
2. Discuss how sample selection relates to π.
3. Discuss a procedure for finding the path π^* in G_S^+ that corresponds to π in N_S^+.
4. Discuss possible cases for π^* in G_S^+ and their relation to ISAC.

[**1. The structure of π.**] We start by assuming that, assuming we are conditioning on Z, there is an unblocked path from D to Y_d in N_S^+. We are going to call this π. We are also able to assume, without loss of generality, that π intersects D only at the starting point of π. What are we able to say about the structure of π across the two halves of N_S^+, the pre-intervention G_S^+ and the post-intervention $(G_S)_{\overline{D}}$? We start by noticing that the elements of Z can only appear on the pre-intervention G_S^+ side of N_S^+. This is because we can only condition on observed variables; we cannot condition on counterfactual variables, which are not observed. This means that we cannot condition on $D = d$ or any of the descendants of $D = d$ in the post-intervention $(G_S)_{\overline{D}}$ side of N_S^+. As such, as soon as π finds it way to the post-intervention side of N_S^+, the remainder of π connecting to Y_d can only contain post-intervention variables, non of which are conditioned on. Moreover, this portion of π in $(G_S)_{\overline{D}}$ can contain only edges pointing toward Y_d. This clarifies that π must be made up of first an unblocked path in the pre-intervention side, G_S^+, that we will label π_1. Next π contains one edge that points from some node in G_S^+ to some node in $(G_S)_{\overline{D}}$, which we label π_2. Recall that we are dealing with graphs in which all bidirected edges have been replaced. We will also see in the next section that π_2 cannot be a bridge. This means that the only type of edge that could connect π_1, which is entirely made up of pre-intervention nodes, to the post-intervention side is a directed edge from the pre-intervention side to the post-intervention side. An edge pointing the other direction would mean that some variables on π_1 are actually post-intervention, a contradiction. Finally, π contains the path we previously discussed, namely, a causal path that contains only descendants of $D = d$ in $(G_S)_{\overline{D}}$ that ends with Y_d. So π is composed of π_1, π_2, π_3. Since N_S^+ is built from G_S^+ and $(G_S)_{\overline{D}}$, π may contain two node "copies" that refer to the same node in G_S^+.

[**2. Sample selection and π.**] How does sample selection relate to π? Sample selection means we condition on $S = 1$. This is a pre-intervention variable. No post-intervention variable can be an ancestor of the pre-intervention version of S, otherwise we would be considering a post-intervention version of S. So all ancestors of the pre-intervention S are also pre-intervention variables. Therefore, all bridges in N_S^+ appear in the pre-intervention side of the graph, G_S^+, since we've assumed that we've replaced bidirected edges with U^*'s with uni-directional edges that point to the nodes that the bidirected edge had pointed to. Hence, any bridge on π will be in π_1. Since we must condition on the pre-intervention S, any path on which the pre-intervention S appears and is not a collider is blocked and so cannot be π. Also, any path on which the pre-intervention S appears and is a collider (or for which S is a descendant of a collider on the path) will correspond to a generalized non-causal path that is identical to the original path except that the collider is not on the generalized non-causal path and the parents of the collider are connected by a bridge on the generalized non-causal path. If the generalized non-causal path is not blocked then the original path will also not be blocked; if the generalized non-causal path is blocked then so is the original path. Therefore, we can limit our analysis to such generalized non-causal paths. So we consider π that do not contain the pre-interventional S, though π may contain bridges in π_1. Since we have assumed that sample selection is not a mediator or a descendant of a mediator, post-intervention versions of S will not appear on π_3 if they exist at all in N_S^+.

[**3. Finding the path π^* in G_S^+ that corresponds to π in N_S^+.**] How can we find a path in G_S^+ that corresponds to π? We follow a procedure laid out in Shpitser et al. (2010). First, we create π', a route in G_S^+, in this way:

1. Start by replacing each instance of a post-intervention variable in π with copy of the same node that appears on the pre-intervention side, G_S^+. We carry along the appropriate occurrence number for each of these replaced nodes.
2. Continue by replacing any instances in which the same variable appears twice in a row with only one copy of that variable. Then reduce the occurrence number of this variable by one and also do this for all the variables that follow.

The portions of π' that were created from π_1 and π_3 (portions of π in N_S^+) will also be unblocked since π_1 and π_3 are unblocked. What about the portion of π created from π_2? This will correspond to a set of three nodes where the center node is the one pointed to by π_2, which we know is a directed edge pointing to some post-intervention node, from the above discussion. The second edge in this triple must be pointing away from the middle node, since all edges in π_3 point toward Y_d, and also must be part of a causal path from D to Y in G_S^+ since the node came from the post-intervention side. But conditioning on nodes on causal paths from D to Y constitutes a violation of ISAC, so we cannot condition on the center node without violating ISAC. Therefore, this last portion of π' is also unblocked, if it exists (it may not if there are no edges in π_3). For example, say that G_S^+ contains $D \to Y$ and $D \to Z \to Y$ and sample selection is not connect to any other node. Then suppose that π is taken to be $D \to Z \leftarrow U_Z \to Z_d \to Y_d$. Here π_1 is $D \to Z \leftarrow U_Z$, π_2 is the edge between U_Z and Z_d, and π_3 is $Z_d \to Y_d$. So π' is $D \to Z \leftarrow U_Z \to Z \to Y$. The node triple in π' that does not correspond to π_1 or π_3 is $U_Z \to Z \to Y$. This is blocked since we condition on Z. However, conditioning on Z is a violation of ISAC since Z lies on a causal path from D to Y in G_S^+. All blocked versions of the node triple in π' that does not correspond to π_1 or π_3 must also violate ISAC for similar reasons. Since the middle node in this node triple is pointed to by π_2, the middle node must be a post-intervention node and so it must lie on a causal path from D to Y in G_S^+, and conditioning on it violates ISAC. Either the node triple is unblocked or it isn't. But, if it isn't, then it could only have resulted from a violation of ISAC. So π' is an unblocked route. By Lemma 2, π^*, the direct route of π' in G_S^+, is an unblocked path in G_S^+.

[4. Possible cases for π^* in G_S^+ and their relation to ISAC.]

So what are the types of π^* we might see and how do these relate to ISAC?

- The easy case is when π^* is a generalized non-causal path. Here we violate ISAC since Z does not block every generalized non-causal path between D and Y that does not pass through S. Note that any time π contains a bridge, π^* will also contain a bridge and π^* will be a generalized non-causal path.
- The case where π^* is a causal path is somewhat more involved. We again assume that π^* is a proper causal path without loss of generality. The first edge in a π that would create

a π^* that is causal would have to be an edge pointing away from some element of D. As we've discussed, π_2 must have been directed in π and pointed to a post-intervention node in $(G_S)_{\overline{D}}$ from a pre-intervention node in G_S^+. The pre-intervention node could not have been a descendant of D, otherwise it would be in the $(G_S)_{\overline{D}}$ part of N_S^+.

- If there are no node copies that are in both π_1 and π_3 (meaning π_1 has a pre-intervention copy and π_3 has a p-intervention copy of the same node), then π^* cannot be a proper causal path from D to Y in G_S^+. The only way it could be would be for the pre-intervention node to be a descendant of D, a contradiction.
- If there are node copies that are in both π_1 and π_3, then the only way to reach the pre-intervention node from D is via a collider unblocked by our conditioning on some element of Z. This would mean that the second node in π (and the second node in π^*) is an ancestor of Z, which violates ISAC.

□

Lemma 4. *If Z d-separates D and Y_d in N_S^+, then $\{Z, S\}$ d-separates D and Y_d in N_S.*

Proof. We very closely follow the structure of the proof of Lemma 3 in the Web Appendix for Daniel et al. (2012). We start by supposing that the statement that "If Z d-separates D and Y_d in N_S^+, then $\{Z, S\}$ d-separates D and Y_d in N_S." is false. This means that, although all paths from D to Y_d in N_S^+ are blocked by Z, we can find a N_S and a Z for which there is a path, ξ, in N_S from D to Y_d that is not blocked by $\{Z, S\}$. The path ξ is also in N_S^+ since N_S^+ is N_S but with edges added. No edges are removed in extending N_S to N_S^+. If ξ is blocked after conditioning on Z in N_S^+ but is unblocked after conditioning on $\{Z, S\}$ in N_S, then either

- There must be a variable on the path ξ that is not in the set $\{Z, S\}$ but the variable is a in Z. In this way, this variable does not block ξ in N_S but does block ξ in N_S^+. But since Z is in $\{Z, S\}$, this is a contradiction.
- There must be a collider on ξ that satisfies both of the following conditions. The

collider is not in Z and does not have descendants in Z. The collider is in $\{Z, S\}$ or has descendants in $\{Z, S\}$. In this way, ξ is blocked in N_S^+ and ξ is not blocked in N_S. Clearly, the collider is S or an ancestor of S. But we can see that ξ is identical to a generalized non-causal path, ξ', in N_S^+ with the exception that the immediate parents of the collider have a bridge between them on ξ' and the collider does not appear on ξ'. If ξ is unblocked in N_S then ξ' must be unblocked in N_S^+. This is because none of the variables along ξ is in $\{Z, S\}$ possibly with the exception of the collider. If they were, then ξ would be blocked. So ξ' must be unblocked in N_S^+, a contradiction.

□

Lemma 5. *If Z d-separates D and Y in G_S^+, then $\{Z, S\}$ d-separates D and Y in G_S.*

Proof. The same argument as in Lemma 4 proves this result. □

Lemma 6. *If $\{Z, S\}$ d-separates D and Y_d in N_S, then $Y_d \perp\!\!\!\perp D|Z, S = 1$ for every model inducing G_S.*

Proof. This follows from Lemma 1. □

Lemma 7. *If $Y_d \perp\!\!\!\perp D|Z, S = 1$ for every model inducing G_S, then $p(Y_d|S = 1) = \sum_z p(Y|D = d, Z = z, S = 1)p(Z = z|S = 1)$, for every model inducing G_S.*

Proof.

$$\begin{aligned}
&p(Y_d|S = 1) \\
&= \sum_z p(Y_d|Z = z, S = 1)p(Z = z|S = 1) && \text{by law of iterated expectations} \\
&= \sum_z p(Y_d|D = d, Z = z, S = 1)p(Z = z|S = 1) && \text{by } Y_d \perp\!\!\!\perp D|Z, S = 1 \\
&= \sum_z p(Y|D = d, Z = z, S = 1)p(Z = z|S = 1) && \text{by consistency}
\end{aligned}$$

□

A.3.2 Selection as a mediator

Lemma 8. *If Z satisfies ISAC in G_S^+ relative to D and Y and S is a mediator between D and Y (but S is not also a descendant of another mediator), then Z d-separates D and $Y_{d,S=1}$ in N_S^+.*

Proof. We rely on the proof of Lemma 3, but with some caveats for the changes due to S being a mediator between D and Y. See that proof for how the objects in this proof are defined. We consider whether there are new types of open paths π that could connect D to $Y_{d,S=1}$ in N_S^+ and if they exist whether they violate ISAC.

[π **that contain a copy of** S**.**] In the extended twin network for $Y_{d,S=1}$, any path along which S appears on the post-intervention side of the graph ($(G_S)_{\overline{D,S}}$) has been severed since we intervene to set $S = 1$ in addition to setting $D = d$. Again π (the assumed unblocked path from D to $Y_{d,S=1}$ in N_S^+) is constructed from three parts: π_1 (an unblocked path in G_S^+), π_3 (a causal path in $(G_S)_{\overline{D,S}}$ on which every node is a descendant of D and/or S), and π_2 (a single edge connecting π_1 and π_3 in N_S^+). This means that π must land in the post-intervention side on a node that is downstream of $S = 1$ and/or of $D = d$. As in Lemma 3, due to the construction of N_S^+ we can focus on the generalized non-causal paths that circumvent the pre-interventional S in G_S^+ as candidates for π_1, rather than any such path that passes through the pre-intervention S. Any path on which the pre-intervention S appears where it is not a collider is blocked. So π is not one of these. The paths on the pre-intervention side on which S is a collider correspond to generalized non-causal paths on which S is circumvented and that are blocked whenever the path on which S is a collider is blocked. Due to our additional intervention on $S = 1$, conditioning on the pre-intervention S, and the bridges that we've added, we can conclude that π will not contain either the pre- or post-interventional copy of S.

[π **that do not contain a copy of** S**.**] Are there any other ways that S as a mediator could add to the cases we need to consider not already covered in Lemma 3? Conditioning on pre-intervention S could open generalized non-causal paths in the pre-intervention side that contain bridges. But if π contains a bridge, then π^* will also contain a bridge and so it will always be non-causal and hence violate ISAC. (Note that we do not allow elements of Z to appear on causal paths or to be descendants of variables on causal paths, whether or not S is on these paths.) Any remaining unblocked π would be of the types already covered by Lemma 3 and the same logic from the proof of Lemma 3 will apply here.

□

Lemma 9. *If Z d-separates D and $Y_{d,S=1}$ in N_S^+, then $\{Z,S\}$ d-separates D and $Y_{d,S=1}$ in N_S.*

Proof. This proof is similar to that for Lemma 4. □

Lemma 10. *If $\{Z,S\}$ d-separates D and $Y_{d,S=1}$ in N_S, then $Y_{d,S=1} \perp\!\!\!\perp D|Z, S=1$ for every model inducing G_S.*

Proof. This follows from Lemma 1. □

Lemma 11. *If $Y_{d,S=1} \perp\!\!\!\perp D|Z, S=1$ for every model inducing G_S, then $p(Y_{D=d,S=1}|S=1) = \sum_z p(Y|D=d, Z=z, S=1)p(Z=z|S=1)$, for every model inducing G_S.*

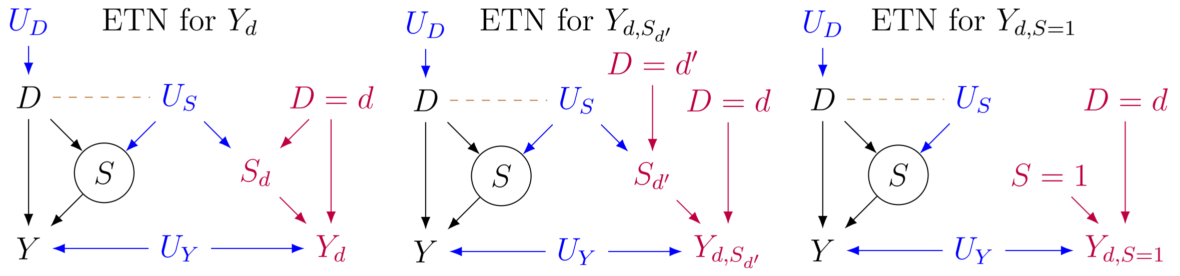
Proof.

$$\begin{aligned}
&p(Y_{d,S=1}|S=1)\\
&\quad= \sum_z p(Y_{d,S=1}|Z=z, S=1)p(Z=z|S=1) && \text{by law of iterated expectations}\\
&\quad= \sum_z p(Y_{d,S=1}|D=d, Z=z, S=1)p(Z=z|S=1) && \text{by } Y_{d,S=1} \perp\!\!\!\perp D|Z, S=1\\
&\quad= \sum_z p(Y|D=d, Z=z, S=1)p(Z=z|S=1) && \text{by consistency}
\end{aligned}$$

□

Lemma 12. *If S is a mediator between D and Y, then no set of observed variables, W, can d-separate D and Y_d (or $Y_{d,S_{d'}}$) in N_S^+, unless S is a deterministic function of observed variables. And so no set of variables, W, (along with S) can d-separate D and Y_d (or $Y_{d,S_{d'}}$) in N_S. Hence, the following do not hold for every model inducing G_S: $Y_d \not\perp\!\!\!\perp D|W, S = 1$ and $Y_{d,S_{d'}} \not\perp\!\!\!\perp D|W, S = 1$.*

Proof. We consider the simplest case in the graphs below. Since we see that we cannot d-separate D and Y_d (or $Y_{d,S_{d'}}$) in N_S^+ in this simplest case, adding more edges and nodes will not change this. The second statement follows from just looking at the twin network contained in the extended twin network here. The last part of the lemma follows from Lemma 1.



□

A.3.3 Selection as a descendant of a mediator

Lemma 13. *If Z satisfies ISAC in G_S^+ relative to D and Y and S is a descendant of a mediator, M, between D and Y (but S is not also a mediator itself), then Z d-separates D and $Y_{d,m}$ in N_S^+.*

Proof. We rely on the proof of Lemma 3 and Lemma 8, but with some caveats for the changes due to S being a descendant of a mediator between D and Y. See those proofs for how the quantities in this proof are defined. Note that we are assuming that S is not itself a mediator between D and Y. We consider whether there are new types of open paths π that could connect D to $Y_{d,S=1}$ in N_S^+ and if they exist whether they violate ISAC.

[π **that contain a copy of** S **or post-intervention** M**.**] Since S is a descendant of M, a mediator between D and Y in G_S^+, but S is not itself a mediator, in the extended twin network for $Y_{d,m}$, N_S^+, any path on which S or M appear on the post-intervention side of the graph $((G_S)_{\overline{D,M}})$ has been severed since we intervene to set $M = m$ and $D = d$, in addition to the fact that S is not a mediator itself. Again, π (the assumed unblocked path from D to $Y_{d,m}$ in N_S^+) is constructed from three parts: π_1 (an unblocked path in G_S^+), π_3 (a causal path in $(G_S)_{\overline{D,M}}$ on which every node is a descendant of D and/or M), and π_2 (a single edge connecting π_1 and π_3 in N_S^+). This means that π must land in the post-intervention side on a node that is downstream of $M = m$ and/or of $D = d$. Further, any path containing a post-intervention version of S will not end with $Y_{d,m}$ since S is not a mediator itself. As in Lemmas 3 and 8, due to the construction of N_S^+ we can focus on the generalized non-causal paths that circumvent S in G_S^+ as candidates for π_1, rather than any such path that passes through S. So π will not contain either the pre- or post-interventional copy of S. It will also not contain a post-interventional copy of M.

[π **that contain pre-intervention** M**.**] Let's consider the cases in which π might have M on the pre-intervention side. This means that any π in N_S^+ between D and $Y_{d,m}$ that contains M must have M in π_1.

- If M is on a non-causal π^* then we have a violation of ISAC.
- As in Lemma 3, if M is on a causal π^* then we will also find a violation of ISAC. π_2, again, must have been directed and pointed from a pre-intervention node to a post-intervention node and the pre-intervention node could not have been a descendant of M, otherwise it would be in the $((G_S)_{\overline{D,M}})$ part of N_S^+. M would be on π_1 and the first edge out of M must be an edge pointing away from M, since M only appears on π_1 and π^* is causal. If there are no node copies that are in both π_1 and π_3 (meaning π_1 has a pre-intervention copy and π_3 has a post-intervention copy of the same node), then π^* cannot be a proper causal path from D to Y in G_S^+. The only way it could

be would be for the pre-intervention node to be a descendant of M, a contradiction. If there are node copies that are in both π_1 and π_3, then the only way to reach the pre-intervention node from M is via a collider unblocked by our conditioning on some element of Z. This would mean that the second node after M in π (and the second node after M in π^*) is an ancestor of Z, which violates ISAC.

So any π on which M appears corresponds to a π^* that violates ISAC.

Due to the above discussion, any generalized non-causal path on which M is an ancestor of Y in G_S^+ does not need to be blocked by Z. This is because such paths cannot correspond to paths from D to $Y_{d,m}$ in N_S^+ unless we condition on a Z that violates the fist condition of ISAC. This is because these paths point to the pre-interventional Y and the post-interventional version of these paths is severed by our intervention setting $M = m$.

Any remaining unblocked π would be of the types already covered by Lemmas 3 and 8 and the same logic from the proofs of Lemmas 3 and 8 will apply here. Note that conditioning on M (that is including it in Z) or variables that are on the same causal path that M is on are prohibited by ISAC.

□

Lemma 14. *If Z d-separates D and $Y_{d,m}$ in N_S^+, then $\{Z, S\}$ d-separates D and $Y_{d,m}$ in N_S.*

Proof. This proof is similar to that for Lemma 4. □

Lemma 15. *If $\{Z, S\}$ d-separates D and $Y_{d,m}$ in N_S, then $Y_{d,m} \perp\!\!\!\perp D|Z, S = 1$ for every model inducing G_S.*

Proof. This follows from Lemma 1. □

Lemma 16. *If Z satisfies ISAC in G_S^+ relative to D and Y and S is a descendant of a mediator, M, between D and Y (but S is not also a mediator itself), then $\{Z, D\}$ d-separates M and $Y_{d,m}$ in N_S^+.*

Proof. We will again show the contrapositive: assuming we are conditioning on Z and D, we show that any unblocked path from M to $Y_{d,m}$ in N_S^+, ϕ, means there is an unblocked path from D to $Y_{d,m}$ in N_S^+ (where we only condition on Z), π, which implies that ISAC is violated in G_S^+ relative to D and Y based on Lemma 13.

We can connect any ϕ that starts with an arrow into M (i.e., $M \leftarrow \ldots Y_{d,m}$ or $M \leftrightarrow \ldots Y_{d,m}$) to $D \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow M$ to create π since M in this case is a collider and we condition on S, a descendant of M; and we cannot condition on any node like W on $D \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow M$ since W would also be on causal paths from D to Y, since M is a mediator, which would be a violation of ISAC. Note that this also holds when M is a direct descendant of D: $D \rightarrow M$ can be connected to ϕ and since M is a collider between D and some variable on ϕ, there is an open π. We can similarly connect any ϕ that starts with a bridge touching M (i.e., $M \text{----} \ldots Y_{d,m}$) can be connected to $D \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow M$ to create π since bridges do not create colliders. Any path with an arrow pointing out of M (i.e., $M \rightarrow \ldots$) can only end at $Y_{d,m}$ after traversing a bridge or a collider that has been conditioned on, since M is a pre-intervention node but $Y_{d,m}$ is a descendant of the intervention $M = m$ not of M. In both of these cases, this path can also be connected with $D \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow M$ to create π. The key is that ϕ is an open path from M to $Y_{d,m}$ and so it can be linked to the causal path from D to M (which cannot be blocked without violating ISAC). Any such unblocked path π from D to $Y_{d,m}$ in N_S^+ (where we only condition on Z) violates ISAC following Lemma 13.

Conditioning on D blocks any open paths between D and $Y_{d,m}$ that could be connected to the causal path from D to M to create an open path between M and $Y_{d,m}$. Could conditioning on D create an open ϕ from M to $Y_{d,m}$ without there also being an open π between D and $Y_{d,m}$? This could only happen if conditioning on D opened a previously closed path, otherwise, no additional paths are opened by conditioning on D (in addition to Z) and we're in the situation above where any ϕ between M and $Y_{d,m}$ can be used to create a π between D and $Y_{d,m}$. Conditioning on D (in addition to Z) would only open paths between M and $Y_{d,m}$ if

D is a collider on such paths. This would mean that there would have to be an open path between D and M that ends with an arrow into D as well as an open path between D and $Y_{d,m}$ that ends with an arrow into D. But any open path between D and $Y_{d,m}$ that ends with an arrow into D is an open π that, as we've seen, would violate ISAC. Note that conditioning on M (that is including it in Z) or variables that are on the same causal path that M is on are prohibited by ISAC.

□

Lemma 17. *If $\{Z, D\}$ d-separates M and $Y_{d,m}$ in N_S^+, then $\{Z, D, S\}$ d-separates M and $Y_{d,m}$ in N_S.*

Proof. This proof is similar to that for Lemma 4. □

Lemma 18. *If $\{Z, D, S\}$ d-separates M and $Y_{d,m}$ in N_S, then $Y_{d,m} \perp\!\!\!\perp M|D, Z, S = 1$ for every model inducing G_S.*

Proof. This follows from Lemma 1. □

Lemma 19. *If $Y_{d,m} \perp\!\!\!\perp D|Z, S = 1$ and $Y_{d,m} \perp\!\!\!\perp M|D, Z, S = 1$ for every model inducing G_S, then $p(Y_{d,m}|S = 1) = \sum_z p(Y|D = d, M = m, Z = z, S = 1)p(Z = z|S = 1)$, for every model inducing G_S.*

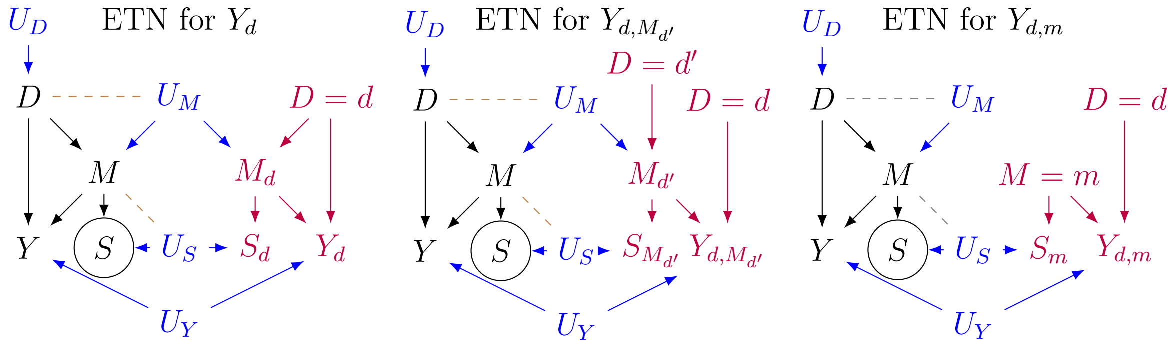
Proof.

$$\begin{aligned}
&p(Y_{d,m}|S = 1)\\
&= \sum_z p(Y_{d,m}|Z = z, S = 1)p(Z = z|S = 1) \text{ by law of iterated expectations}\\
&= \sum_z p(Y_{d,m}|D = d, Z = z, S = 1)p(Z = z|S = 1) \text{ by } Y_{d,m} \perp\!\!\!\perp D|Z, S = 1\\
&= \sum_z p(Y_{d,m}|D = d, M = m, Z = z, S = 1)p(Z = z|S = 1) \text{ by } Y_{d,m} \perp\!\!\!\perp M|D, Z, S = 1\\
&= \sum_z p(Y|D = d, M = m, Z = z, S = 1)p(Z = z|S = 1) \text{ by consistency}
\end{aligned}$$

□

Lemma 20. *If S is a descendant of a mediator, M, between D and Y, then no set of observed variables, W, can d-separate D and Y_d (or $Y_{d,M_{d'}}$) in N_S^+, unless S is a deterministic function of observed variables. And so no set of variables, W, (along with S) can d-separate D and Y_d (or $Y_{d,M_{d'}}$) in N_S. Hence, the following do not hold for every model inducing G_S: $Y_d \not\perp\!\!\!\perp D|W, S = 1$ and $Y_{d,S_{d'}} \not\perp\!\!\!\perp D|W, S = 1$.*

Proof. We consider the simplest case in the graphs below. Since we see that we cannot d-separate D and Y_d (or $Y_{d,M_{d'}}$) in N_S^+ in this simplest case, adding more edges and nodes will not change this. The second statement follows from just looking at the twin network contained in the extended twin network here. The last part of the lemma follows from Lemma 1.



□

A.3.4 Generalization

Lemma 21. *If a set of nodes Z in G_S^+ satisfies GC relative to D (treatment) and Y (outcome) and S is not a mediator or descendant of a mediator between D and Y, then Z_{Ext} d-separates S and Y_d in N_S^+.*

Proof. We take a similar approach as in the proof of Lemma 3. We will show the contrapositive: assuming that we are conditioning on Z, which includes Z_{Ext}, an unblocked path from S to

Y_d in N_S^+ implies that GC is violated in G_S^+ relative to D and Y.

- We define π like in Lemma 3: We start by assuming that, assuming we are conditioning on Z, there is an unblocked path, π, from S to Y_d in N_S^+. Like in Lemma 3, we can consider π that do not contain the pre-interventional S, though π may contain bridges in π_1. Since we have assumed that sample selection is not a mediator or a descendant of a mediator, post-intervention versions of S will not appear on π_3 if they exist at all in N_S^+.
- As in in Lemma 3, we can create a π' that is an unblocked route in G_S^+ and π^*, the direct route of π' in G_S^+, is an unblocked path in G_S^+.
- Now, let's consider the types of paths that π^* could be.
  - π^* that do not end in a causal path from D to Y violate of GC.
  - For π^* that do end in a causal paths from D to Y, we want to show that these would imply contradictions for π or violate GC.
    - How can we have D on π^*? This could only result from the pre-intervention side copy of D appearing on π, since no edges point into $D = d$, no bridges can connect to $D = d$, and π must land downstream of $D = d$ on the post-intervention side.
    - So how can the pre-intervention side copy of D appear on π? This could only result from a non-causal or causal path from S to D, where both are on the pre-intervention side. So this non-causal or causal path from S to D is entirely on the pre-intervention side and, therefore, is part of π_1.
    - Then the question becomes: how can we get a causal path from D to Y in π^*, when D is pre-intervention and Y_d is post-intervention on π? From here we can follow the logic in the section bullet of part 4. of the proof of Lemma 3, which shows us that this either results in a contradiction or violates GC since it violates ISAC.

□

Lemma 22. *If Z_{Ext} d-separates S and Y_d in N_S^+, then Z_{Ext} d-separates S and Y_d in N_S.*

Proof. We start by supposing that the statement "If Z_{Ext} d-separates S and Y_d in N_S^+, then Z_{Ext} d-separates S and Y_d in N_S." is false. This means that although all paths from S to Y_d in N_S^+ are blocked by Z_{Ext}, we can find a N_S and a Z_{Ext} for which there is a path, ξ, in N_S from S to Y_d that is not blocked by Z_{Ext}. Without loss of generality, we assume that ξ only intersects S at the endpoint. N_S^+ is identical to N_S except for N_S^+ contains bridges created as a result of conditioning on S. Bridges are added to N_S between all variables between which S is a collider or an ancestor of S is a collider resulting in N_S^+. If ξ is unblocked in N_S, then ξ traverses no bridges, since bridges do not appear in N_S. The path ξ is also in N_S^+ and is unblocked since N_S^+ is N_S but with edges added and we are conditioning on the same set of nodes, Z_{Ext}, in both. No edges are removed in extending N_S to N_S^+. In this way, N_S^+ "contains" all of N_S. But an unblocked path between S and Y_d in N_S^+ is a contradiction. □

Lemma 23. *If Z_{Ext} d-separates S and Y_d in N_S, then $Y_d \perp\!\!\!\perp S|Z_{Ext}$ for every model inducing G_S.*

Proof. This follows from Lemma 1. □

Lemma 24. *If $Z_{Ext} \subset Z$ and $Z_{Int} = Z - Z_{Ext}$, $Y_d \perp\!\!\!\perp S|Z_{Ext}$, and $Y_d \perp\!\!\!\perp D|Z, S = 1$ for every model inducing G_S, then, for every model inducing G_S,*

$$\begin{aligned}
&p(Y_d|S=1)\\
&= \sum_z p(Y_d|Z=z, S=1)p(Z_{Int}=z_{Int}|Z_{Ext}=z_{Ext}, S=1)p(Z_{Ext}=z_{Ext}|S=1)\\
&= \sum_z p(Y|D=d, Z=z, S=1)p(Z_{Int}=z_{Int}|Z_{Ext}=z_{Ext}, S=1)p(Z_{Ext}=z_{Ext}|S=1)\\
&\textit{and } p(Y_d)\\
&= \sum_z p(Y_d|Z=z, S=1)p(Z_{Int}=z_{Int}|Z_{Ext}=z_{Ext}, S=1)p(Z_{Ext}=z_{Ext})\\
&= \sum_z p(Y|D=d, Z=z, S=1)p(Z_{Int}=z_{Int}|Z_{Ext}=z_{Ext}, S=1)p(Z_{Ext}=z_{Ext}).
\end{aligned}$$

Proof.

$$
\begin{aligned}
&p(Y_d|S=1)\\
&=\sum_z p(Y_d|Z=z,S=1)p(Z=z|S=1)\\
&=\sum_z p(Y_d|D=d,Z=z,S=1)p(Z=z|S=1) \text{ by } Y_d \perp\!\!\!\perp D|Z,S=1\\
&=\sum_z p(Y|D=d,Z=z,S=1)p(Z=z|S=1) \text{ by consistency}\\
&=\sum_z p(Y|D=d,Z=z,S=1)p(Z_{\text{Int}}=z_{\text{Int}}|Z_{\text{Ext}}=z_{\text{Ext}},S=1)p(Z_{\text{Ext}}=z_{\text{Ext}}|S=1)
\end{aligned}
$$

$$
\begin{aligned}
&p(Y_d)\\
&=\sum_{z_{\text{Ext}}} p(Y_d|Z_{\text{Ext}}=z_{\text{Ext}})p(Z_{\text{Ext}}=z_{\text{Ext}})\\
&=\sum_{z_{\text{Ext}}} p(Y_d|Z_{\text{Ext}}=z_{\text{Ext}},S=1)p(Z_{\text{Ext}}=z_{\text{Ext}}) \text{ by } Y_d \perp\!\!\!\perp S|Z_{\text{Ext}}\\
&=\sum_{z_{\text{Ext}}} \left[\sum_{z_{\text{Int}}} p(Y_d,Z_{\text{Int}}=z_{\text{Int}}|Z_{\text{Ext}}=z_{\text{Ext}},S=1)\right] p(Z_{\text{Ext}}=z_{\text{Ext}})\\
&=\sum_{z_{\text{Ext}}} \left[\sum_{z_{\text{Int}}} p(Y_d|Z_{\text{Int}}=z_{\text{Int}},Z_{\text{Ext}}=z_{\text{Ext}},S=1)p(Z_{\text{Int}}=z_{\text{Int}}|Z_{\text{Ext}}=z_{\text{Ext}},S=1)\right] p(Z_{\text{Ext}}=z_{\text{Ext}})\\
&=\sum_z p(Y_d|Z=z,S=1)p(Z_{\text{Int}}=z_{\text{Int}}|Z_{\text{Ext}}=z_{\text{Ext}},S=1)p(Z_{\text{Ext}}=z_{\text{Ext}})\\
&=\sum_z p(Y_d|D=d,Z=z,S=1)p(Z_{\text{Int}}=z_{\text{Int}}|Z_{\text{Ext}}=z_{\text{Ext}},S=1)p(Z_{\text{Ext}}=z_{\text{Ext}})\\
&\text{by } Y_d \perp\!\!\!\perp D|Z,S=1\\
&=\sum_z p(Y|D=d,Z=z,S=1)p(Z_{\text{Int}}=z_{\text{Int}}|Z_{\text{Ext}}=z_{\text{Ext}},S=1)p(Z_{\text{Ext}}=z_{\text{Ext}}) \text{ by consistency}
\end{aligned}
$$

□

A.4 Theorems

Theorem 6. *If a set of nodes Z in internal selection graph G_S^+ satisfies ISAC relative to D (treatment) and Y (outcome) and S is not a mediator or descendant of a mediator between D and Y, then, for every model inducing G_S, $Y_d \perp\!\!\!\perp D|Z, S = 1$ and we can then identify $p(Y_d|S = 1) = \sum_z p(Y|d, z, S = 1)p(z|S = 1)$, all of which is estimable from the selected sample alone.*

Proof. Lemmas 3, 4, 6, and 7 prove the result. □

Theorem 7. *If a set of nodes Z in internal selection graph G_S^+ satisfies ISAC relative to D (treatment) and Y (outcome) and S is a mediator between D and Y (but S is not also a descendant of another mediator), then, for every model inducing G_S, $Y_{d,S=1} \perp\!\!\!\perp D|Z, S = 1$ and we can then identify $p(Y_{d,S=1}|S = 1) = \sum_z p(Y|d, z, S = 1)p(z|S = 1)$, all of which is estimable from the selected sample alone. Further note that, for any set of observables, W, $Y_d \not\perp\!\!\!\perp D|W, S = 1$ and $Y_{d,S_{d'}} \not\perp\!\!\!\perp D|W, S = 1$, unless S is a deterministic function of observed variables.*

Proof. Lemmas 8, 9, 10, 11, and 12 prove the result. □

Theorem 8. *If a set of nodes Z in internal selection graph G_S^+ satisfies ISAC to D (treatment) and Y (outcome), where S is a descendant of an observed mediator, M, between D and Y (but S is not also a mediator itself), then, for every model inducing G_S, $Y_{d,m} \perp\!\!\!\perp D|Z, S = 1$ and $Y_{d,m} \perp\!\!\!\perp M|D, Z, S = 1$, where $M = m$ is a value observed in the sample. We can then identify, for every model inducing G_S, $p(Y_{d,m}|S = 1) = \sum_z p(Y|d, m, z, S = 1)p(z|S = 1)$, all of which is estimable from the selected sample alone. Further note that, for any set of observables, W, $Y_d \not\perp\!\!\!\perp D|W, S = 1$ and $Y_{d,M_{d'}} \not\perp\!\!\!\perp D|W, S = 1$, unless S is a deterministic function of observed variables.*

Proof. Lemmas 13, 14, 15, 16, 17, 18, 19, and 20 prove the result. □

If S is both a mediator and a descendant of a mediator (M) between D and Y, then we might consider potential outcomes of the form $Y_{d,m,S=1}$ (intervening to set $D = d$, $M = m$, and $S = 1$) and something like these theorems should hold. We do not demonstrate this for brevity. The following result concerning generalization translates the results in Correa et al. (2018) to use potential outcomes.

Theorem 9. *If a set of nodes Z in G_S^+ satisfies GC relative to D (treatment) and Y (outcome) and S is not a mediator or descendant of a mediator between D and Y, then $Y_d \perp\!\!\!\perp D|Z, S = 1$ and $Y_d \perp\!\!\!\perp S|Z_{Ext}$. We can identify $p(Y_d) = \sum_z p(Y|D = d, Z = z, S = 1)p(Z_{Int} = z_{Int}|Z_{Ext} = z_{Ext}, S = 1)p(Z_{Ext} = z_{Ext})$, where $Z_{Int} = Z - Z_{Ext}$.*

Proof. Theorem 6 and Lemmas 21, 22, 23, and 24 prove the result. □

A.5 IPW estimation

Following related discussions in Hernán and Robins (2006); VanderWeele (2009); Correa et al. (2018); Hernán and Robins (2020) and elsewhere, we present IPW estimators for $\mathbb{E}[Y_d|S = 1]$, $\mathbb{E}[Y_{d,S=1}|S = 1]$, and $\mathbb{E}[Y_{d,m}|S = 1]$. These are familiar results but tailored to our internal validity for the selected sample focus.

If $Y_d \perp\!\!\!\perp D|Z, S = 1$ (following an application of ISAC) and we have SUTVA, consistency, and positivity, then we can show that, by the law of large numbers, the IPW estimator

$$\hat{\mu}_d = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i \times \mathbb{1}_{D_i=d}}{\hat{p}(D_i = d|Z_i = z, S = 1)}$$

is consistent for $\mathbb{E}[Y_d|S = 1]$ when the propensity score model $\hat{p}(D_i = d|Z_i = z, S = 1)$ is correctly specified:

$$\begin{aligned}
\mathbb{E}[Y_d|S=1] &= \sum_y y \times p(Y_d = y|S=1) \\
&= \sum_y \sum_z y \times p(Y=y|D=d, Z=z, S=1) p(Z=z|S=1) \quad \text{by Lemma 7} \\
&= \sum_y \sum_z y \times \frac{p(Y=y, D=d|Z=z, S=1)}{p(D=d|Z=z, S=1)} p(Z=z|S=1) \\
&= \sum_y \sum_z y \times \frac{p(Y=y, D=d, Z=z|S=1)}{p(D=d|Z=z, S=1)} \\
&= \sum_y \sum_z \sum_d y \times \mathbb{1}_{D=d} \times \frac{p(Y=y, D=d, Z=z|S=1)}{p(D=d|Z=z, S=1)} \\
&= \mathbb{E}\left[\frac{Y \times \mathbb{1}_{D=d}}{p(D=d|Z=z, S=1)} | S=1\right] = \mathbb{E}[\hat{\mu}_d|S=1]
\end{aligned}$$

If $Y_{d,S=1} \perp\!\!\!\perp D|Z, S=1$ and we have SUTVA, consistency, and positivity, then we can show that, by the law of large numbers, the IPW estimator $\hat{\mu}_d$ is consistent for $\mathbb{E}[Y_{d,S=1}|S=1]$ when the propensity score model $\hat{p}(D=d|Z=z, S=1)$ is correctly specified:

$$\begin{aligned}
\mathbb{E}[Y_{d,S=1}|S=1] &= \sum_y y \times p(Y_{d,S=1} = y|S=1) \\
&= \sum_y \sum_z y \times p(Y=y|D=d, Z=z, S=1) p(Z=z|S=1) \quad \text{by Lemma 11} \\
&= \sum_y \sum_z y \times \frac{p(Y=y, D=d|Z=z, S=1)}{p(D=d|Z=z, S=1)} p(Z=z|S=1) \\
&= \sum_y \sum_z y \times \frac{p(Y=y, D=d, Z=z|S=1)}{p(D=d|Z=z, S=1)} \\
&= \sum_y \sum_z \sum_d y \times \mathbb{1}_{D=d} \times \frac{p(Y=y, D=d, Z=z|S=1)}{p(D=d|Z=z, S=1)} \\
&= \mathbb{E}\left[\frac{Y \times \mathbb{1}_{D=d}}{p(D=d|Z=z, S=1)} | S=1\right] = \mathbb{E}[\hat{\mu}_d|S=1]
\end{aligned}$$

If $Y_{d,m} \perp\!\!\!\perp D|Z, S = 1$ and $Y_{d,m} \perp\!\!\!\perp M|D, Z, S = 1$ and we have SUTVA, consistency, and positivity, then we can show that, by the law of large numbers, the IPW estimator

$$\hat{\mu}_{d,m} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \times \mathbb{1}_{M_i=m} \times \mathbb{1}_{D_i=d}}{\hat{p}(M_i = m|D_i = d, Z_i = z, S = 1)\hat{p}(D_i = d|Z_i = z, S = 1)}$$

is consistent for $\mathbb{E}[Y_{d,m}|S = 1]$ when the propensity score model $\hat{p}(D_i = d|Z_i = z, S = 1)$ and the mediator model $\hat{p}(M_i = m|D_i = d, Z_i = z, S = 1)$ are correctly specified:

$$\begin{aligned}
&\mathbb{E}[Y_{d,m}|S = 1] \\
&= \sum_y y \times p(Y_{d,m} = y|S = 1) \\
&= \sum_y \sum_z y \times p(Y = y|D = d, M = m, Z = z, S = 1)p(Z = z|S = 1) \text{ by Lemma 19} \\
&= \sum_y \sum_z y \times \frac{p(Y = y, M = m|D = d, Z = z, S = 1)}{p(M = m|D = d, Z = z, S = 1)} p(Z = z|S = 1) \\
&= \sum_y \sum_z y \times \frac{p(Y = y, M = m|D = d, Z = z, S = 1)}{p(M = m|D = d, Z = z, S = 1)} \frac{p(D = d|Z = z, S = 1)}{p(D = d|Z = z, S = 1)} p(Z = z|S = 1) \\
&= \sum_y \sum_z y \times \frac{p(Y = y, M = m, D = d|Z = z, S = 1)}{p(M = m|D = d, Z = z, S = 1)p(D = d|Z = z, S = 1)} p(Z = z|S = 1) \\
&= \sum_y \sum_z y \times \frac{p(Y = y, M = m, D = d, Z = z|S = 1)}{p(M = m|D = d, Z = z, S = 1)p(D = d|Z = z, S = 1)} \\
&= \sum_y \sum_z \sum_m \sum_d y \times \mathbb{1}_{M=m} \times \mathbb{1}_{D=d} \times \frac{p(Y = y, M = m, D = d, Z = z|S = 1)}{p(M = m|D = d, Z = z, S = 1)p(D = d|Z = z, S = 1)} \\
&= \mathbb{E}\left[\frac{Y \times \mathbb{1}_{M=m} \times \mathbb{1}_{D=d}}{p(M = m|D = d, Z = z, S = 1)p(D = d|Z = z, S = 1)} |S = 1\right] \\
&= \mathbb{E}[\hat{\mu}_{d,m}|S = 1]
\end{aligned}$$

A.6 Conditioning on a collider

Shahar and Shahar (2017) discuss the conditions under which an association is created between the parents of a collider when the collider is conditioned on for discrete variables.

They show that, “[i]f [D] and [U] are marginally independent causes of [S], then [D] and [U] are dependent conditional on [$S = 1$] if and only if [D] and [U] modify each other’s effects on [$S = 1$].” These authors define effects on the collider as well as effect modification in terms of probability ratios. Section A.7 of Appendix A further shows that non-zero interaction information is a requirement for dependency to be created between two marginally independent parents of a collider, when the collider is conditioned on. Also relevant to the discussion in the present paper, Shahar and Shahar (2017) show that for marginally independent causes D, U of a binary collider S, if the effects of D and U on S are not null, then D, U modify each other’s effects in at least one stratum of S and are dependent in at least one stratum of S.

Using our working example, let us consider what would be required for the parents of the sample selection node to remain independent after stratifying to $S = 1$. Simplifying the example inspired by Knox et al. (2020), we consider how the effect of police perception of majority vs minority race ($D \in \{\text{majority}, \text{minority}\}$) on police making a stop or not ($S \in \{0, 1\}$) might be modified by an unobserved factor that represents police officer stress levels ($U \in \{\text{high}, \text{low}\}$). Following Shahar and Shahar (2017), we consider the probability ratios in (A.1):

$$\frac{p(S = 1|D = \text{majority}, U = \text{high})}{p(S = 1|D = \text{minority}, U = \text{high})} \overset{?}{=} \frac{p(S = 1|D = \text{majority}, U = \text{low})}{p(S = 1|D = \text{minority}, U = \text{low})} \tag{A.1}$$

The left hand side of (A.1) is the effect of police perception of race on police making a stop when police officers are under high stress. The right hand side of (A.1) is the effect of police perception of race on police making a stop when police officers are under low stress. Police officer stress not modifying the effect of police perception of race on police making a stop would mean that the left and right hand sides of (A.1) are equal. Under what circumstances would this occur?

The simplest scenario would be when people perceived to be from the majority racial group are never stopped. That is, the numerators on both sides of (A.1) are zero: $p(S = 1|D = \text{majority}, U = \text{high}) = 0$ and $p(S = 1|D = \text{majority}, U = \text{low}) = 0$. This would make both probability ratios zero, meaning no effect moderation. This is clearly an absurd scenario. The other case would be when the two probability ratios perfectly balance, meaning that there is no effect moderation. This is also not likely; police officer stress levels likely do modify the effect of perceptions of race on making a stop in some way. But if one of these scenarios holds, following Shahar and Shahar (2017)'s result quoted above it can be shown that $p(D = \text{majority}|U = \text{high}, S = 1) = p(D = \text{majority}|S = 1)$. We demonstrate this in full below. That is, police perception of race is independent of police stress levels, despite stratifying to $S = 1$, if one the the two (unrealistic) scenarios holds. Further, if one the the two scenarios holds and so long as $p(S = 1|D = \text{minority}, U = \text{high}) \neq p(S = 1|D = \text{minority}, U = \text{low})$ (i.e., there is an effect from police officer stress), then we would see that the two sides of (A.2) are not equal, meaning there is effect moderation and hence police perception of race is dependent on police stress levels, stratifying to $S = 0$.

$$\frac{p(S = 0|D = \text{majority}, U = \text{high})}{p(S = 0|D = \text{minority}, U = \text{high})} \overset{?}{=} \frac{p(S = 0|D = \text{majority}, U = \text{low})}{p(S = 0|D = \text{minority}, U = \text{low})} \tag{A.2}$$

Practitioners would only know that they're in the setting in which stratifying to $S = 1$ does not create association between the parents of S if they have detailed knowledge of the selection mechanism, like the unrealistic scenarios above. In such a situation, even if the parents of the sample selection node are dependent due to selection, the practitioner could use inverse probability of selection weighting to estimate unbiased effects. (Thompson and Arah, 2014) But we emphasize that such knowledge is hard to come by and the assumptions required to fall into such a setting will often be absurd.

We follow Shahar and Shahar (2017) to show that $p(D = \text{majority}|U = \text{high}, S = 1) = p(D = \text{majority}|S = 1)$. In the derivation below, we abbreviate "majority" as "ma" and "minority" as "mi." We rely on the fact that there is no effect moderation to show this. First note that we can write

$$\frac{p(S=1|D=\text{ma},U=\text{high})}{p(S=1|D=\text{mi},U=\text{high})} = \frac{p(S=1|D=\text{ma},U=\text{high})p(U=\text{high}|D=\text{ma})}{p(S=1|D=\text{mi},U=\text{high})p(U=\text{high}|D=\text{mi})}$$
$$= \frac{p(S=1,U=\text{high}|D=\text{ma})}{p(S=1,U=\text{high}|D=\text{mi})}$$

Next we show that we can write

$$\frac{p(S=1|D=\text{ma},U=\text{high})}{p(S=1|D=\text{mi},U=\text{high})}$$
$$= \frac{p(S=1|D=\text{ma},U=\text{high})}{p(S=1|D=\text{mi},U=\text{high})}\frac{p(S=1,U=\text{high}|D=\text{mi})}{p(S=1|D=\text{mi})}$$
$$+ \frac{p(S=1|D=\text{ma},U=\text{low})}{p(S=1|D=\text{mi},U=\text{low})}\frac{p(S=1,U=\text{low}|D=\text{mi})}{p(S=1|D=\text{mi})}$$
$$= \frac{p(S=1,U=\text{high}|D=\text{ma})}{p(S=1,U=\text{high}|D=\text{mi})}\frac{p(S=1,U=\text{high}|D=\text{mi})}{p(S=1|D=\text{mi})}$$
$$+ \frac{p(S=1,U=\text{low}|D=\text{ma})}{p(S=1,U=\text{low}|D=\text{mi})}\frac{p(S=1,U=\text{low}|D=\text{mi})}{p(S=1|D=\text{mi})}$$
$$= \frac{p(S=1,U=\text{high}|D=\text{ma})}{p(S=1|D=\text{mi})} + \frac{p(S=1,U=\text{low}|D=\text{ma})}{p(S=1|D=\text{mi})} = \frac{p(S=1|D=\text{ma})}{p(S=1|D=\text{mi})}$$

Finally, we show that

$$
\begin{aligned}
&p(D = \text{majority}|U = \text{high}, S = 1)\\
&= \frac{p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma}|U = \text{high})}{p(S = 1|U = \text{high})}\\
&= \frac{p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma}|U = \text{high})}{\begin{pmatrix} p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma}|U = \text{high}) + \\ p(S = 1|D = \text{mi}, U = \text{high})p(D = \text{mi}|U = \text{high}) \end{pmatrix}}\\
&= \frac{p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma})}{p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma}) + p(S = 1|D = \text{mi}, U = \text{high})p(D = \text{mi})}\\
&= p(D = \text{ma}) \left[\begin{pmatrix} \frac{p(S=1|D=\text{ma},U=\text{high})}{p(S=1|D=\text{ma},U=\text{high})} p(D = \text{ma}) + \\ \frac{p(S=1|D=\text{ma},U=\text{high})}{p(S=1|D=\text{mi},U=\text{high})} p(D = \text{mi}) \end{pmatrix} \right]^{-1}\\
&= p(D = \text{ma}) \left[\frac{p(S = 1|D = \text{ma})}{p(S = 1|D = \text{ma})} p(D = \text{ma}) + \frac{p(S = 1|D = \text{ma})}{p(S = 1|D = \text{mi})} p(D = \text{mi}) \right]^{-1} \quad \text{from above}\\
&= \frac{p(S = 1|D = \text{ma})p(D = \text{ma})}{p(S = 1|D = \text{ma})p(D = \text{ma}) + p(S = 1|D = \text{mi})p(D = \text{mi})}\\
&= \frac{p(S = 1|D = \text{ma})p(D = \text{ma})}{p(S = 1)}\\
&= p(D = \text{majority}|S = 1)
\end{aligned}
$$

A.7 Colliders and interaction information

We draw on discussion of interaction information and colliders from Ghassami and Kiyavash (2017) (also see McGill (1954)) to show that zero interaction information for marginally independent random variables implies continued independence conditional on a collider. Section A.6 of Appendix A further discusses collider stratification in the context of our working example. Suppose we have two random variables X_1, X_2 and variable S that is a collider between X_1, X_2: $X_1 \to S \leftarrow X_2$. Interaction information is a generalization of mutual information to the case when there are multiple variables. See Cover and Thomas (2006) for an introduction to mutual information. We can write interaction information as

(Ghassami and Kiyavash, 2017)

$$\begin{aligned}\mathrm{MI}(S;X_1;X_2) &= \mathrm{MI}(X_1;X_2) - \mathrm{MI}(X_1;X_2|S)\\ &= \mathrm{MI}(S;X_1) - \mathrm{MI}(S;X_1|X_2)\\ &= \mathrm{MI}(S;X_2) - \mathrm{MI}(S;X_2|X_1)\end{aligned}$$

Using the first expression for interaction information above, we see that $\mathrm{MI}(X_1;X_2|S) = \mathrm{MI}(X_1;X_2) - \mathrm{MI}(S;X_1;X_2)$. It is trivial to see that, $\mathrm{MI}(S;X_1;X_2) = 0$ means that $\mathrm{MI}(X_1;X_2|S) = \mathrm{MI}(X_1;X_2)$. Meaning that zero interaction information implies that there is no change in mutual information between X_1, X_2 when we condition on the collider between them. For marginally independent X_1, X_2, $\mathrm{MI}(X_1;X_2|S) = -\mathrm{MI}(S;X_1;X_2)$ and so if $\mathrm{MI}(S;X_1;X_2) = 0 \implies \mathrm{MI}(X_1;X_2|S) = 0$. Meaning that zero interaction information implies that conditional on the collider, X_1, X_2 remain independent.

Another view of interaction information is possible by writing

$$\begin{aligned}\mathrm{MI}(S;[X_1,X_2]) &= \mathrm{MI}(S;X_2) + \mathrm{MI}(S;X_1|X_2)\\ &\implies \mathrm{MI}(S;[X_1,X_2]) - \mathrm{MI}(X_1;S) - \mathrm{MI}(X_2;S)\\ &\qquad= \mathrm{MI}(S;X_2) + \mathrm{MI}(S;X_1|X_2) - \mathrm{MI}(S;X_1) - \mathrm{MI}(S;X_2)\\ &\qquad= -(\mathrm{MI}(S;X_1) - \mathrm{MI}(S;X_1|X_2))\\ &\qquad= -\mathrm{MI}(S;X_1;X_2)\\ \implies \mathrm{MI}(S;X_1;X_2) &= \mathrm{MI}(X_1;S) + \mathrm{MI}(X_2;S) - \mathrm{MI}(S;[X_1,X_2])\end{aligned}$$

where $\mathrm{MI}(S;[X_1,X_2]) = \mathrm{MI}(S;X_2) + \mathrm{MI}(S;X_1|X_2)$ captures the information that X_1, X_2 jointly share with S. Interaction information would be zero when $\mathrm{MI}(X_1;S) + \mathrm{MI}(X_2;S)$ equals $\mathrm{MI}(S;[X_1,X_2])$. That is, when the information that X_1, X_2 jointly share with S is exactly equal to the the mutual information between X_1 and S added to the mutual information between X_2 and S. This would arise when the information shared between X_1 and S does not overlap with the information shared between X_2 and S.

We include this discussion only to show that conditioning on a collider does not necessarily lead to mutual information (and therefore dependence) between marginally independent parents of the collider or a change in the mutual information between marginally dependent parents of the collider. We've shown that it does not precisely when interaction information is zero.

A few additional notes on interaction information. If X_1 and X_2 are marginally independent causes of S, one might be tempted to think of interaction information as arising through the "interaction" between X_1 and X_2 in determining S, since no shared information existed marginally. However, interaction information can be difficult to interpret and what is meant by this "interaction" is not necessarily what we might expect. See a Krippendorff (2009) for a discussion of these difficulties, including that interaction information can be negative.

The "interaction" is not necessarily something like an interaction term in a linear model. It is easy to show that for Gaussian random variables $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0,1)$, and $S = X1 + X2 + \epsilon$, where there is no interaction term $X_1 \times X_2$ in the data generating process for S, there is still non-zero interaction information between X_1, X_2, S. This arises from the simple fact that mutual information for Gaussians has the following relationship with R^2's: $\text{MI}(A;B) = -\frac{1}{2}\log(1 - R^2_{A,B})$. (Cover and Thomas, 2006) From above, we know that $\text{MI}(S;X_1;X_2) = \text{MI}(X_1;S) + \text{MI}(X_2;S) - \text{MI}(S;[X_1,X_2])$. From the relationship between mutual information and R^2, we have that $\text{MI}(X_1;S) = -\frac{1}{2}\log(1 - R^2_{X_1,S}) = -\frac{1}{2}\log(1-\frac{1}{3}) \approx 0.203$ and similarly, $\text{MI}(X_2;S) \approx 0.203$. However $\text{MI}(S;[X_1,X_2]) = -\frac{1}{2}\log(1 - R^2_{S,X_1+X_2}) = -\frac{1}{2}\log(1-\frac{2}{3}) \approx 0.55 \neq 0.406 \approx \text{MI}(X_1;S)+\text{MI}(X_2;S)$. So we see that interaction information is non-zero, this in turn means conditional mutual information is non-zero, i.e. $\text{MI}(X_1;X_2|S) \neq 0$. The partial R^2, $R^2_{X_1,X_2|S} = \left(\frac{R_{X_1,X_2} - R_{X_1,S}R_{X_2,S}}{\sqrt{1-R^2_{X_1,S}}\sqrt{1-R^2_{X_2,S}}}\right)^2 = \frac{R^2_{X_1,S}R^2_{X_2,S}}{(1-R^2_{X_1,S})(1-R^2_{X_2,S})} = \frac{(\frac{1}{3})^2}{(1-\frac{1}{3})^2} = \frac{1}{4}$, is also not zero.

We note two additional interesting cases. First, even if $\text{MI}(X_1;S)$ and $R^2_{X_1,S}$ are zero, we can have non-zero interaction information, conditional mutual information and partial R^2 if the zeroes are due to perfect balancing of the path coefficients (e.g., $X_1 \sim \mathcal{N}(0,1)$,

$\xi \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0,1)$, $X_2 = X_1 + \xi$ and $S = -X_1 + X_2 + \epsilon$). Second, even when $R^2_{X_1,X_2|S}$ and $\mathrm{MI}(X_1;X_2|S)$ are zero, we can have non-zero interaction information. The zero $R^2_{X_1,X_2|S}$ could be due to perfect balancing between non-zero R_{X_1,X_2} and R_{S,X_1}, and R_{S,X_2}, which coincides with zero $\mathrm{MI}(X_1;X_2|S)$ due to perfect balancing between non-zero mutual information $\mathrm{MI}(X_1;X_2)$ and interaction information $\mathrm{MI}(S;X_1;X_2)$. For example, if the data generating process is $X_1 \sim \mathcal{N}(0,1)$, $\xi \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0,1)$, $X_2 = \gamma X_1 + \xi$ and $S = \sqrt{\gamma} X_1 + \sqrt{\gamma} X_2 + \epsilon$. In this example the marginal dependency between X_1 and X_2 is eliminated when we condition on S.

A.8 Exercises

Here we consider additional substantive examples that highlight different ways internal validity can relate to sample selection.

A.8.1 Exercise: Height and Free Throw Percentage in the NBA

Consider an oddity of the sporting world: taller players tend to have lower free throw percentages in the NBA. (McMahan, 2017; Helin, 2011) Looking at NBA player-season free-throw percentages from 1950 - 2017 versus player height in Figure A.4, we see a negative relationship with a correlation of -0.25.[4] Indeed, a simple linear regression produces a slope estimate of a decrease of 0.3 percentage points in free throw percentage for each additional centimeter of height; see Table A.1. This estimate is meant only as a statistical summary not a causal effect, indeed the relationship appears to be non-linear.

But what might the underlying causal relationships look like? Does being taller cause you to make fewer free throws? Likely not. Perhaps there is a slightly positive direct effect of height on shooting, since taller people are just closer to the basket. Additionally, taller

[4] These data come from https://www.basketball-reference.com/ and https://www.kaggle.com/datasets/drgilermo/nba-players-stats.

Figure A.4: Free Throws and Height

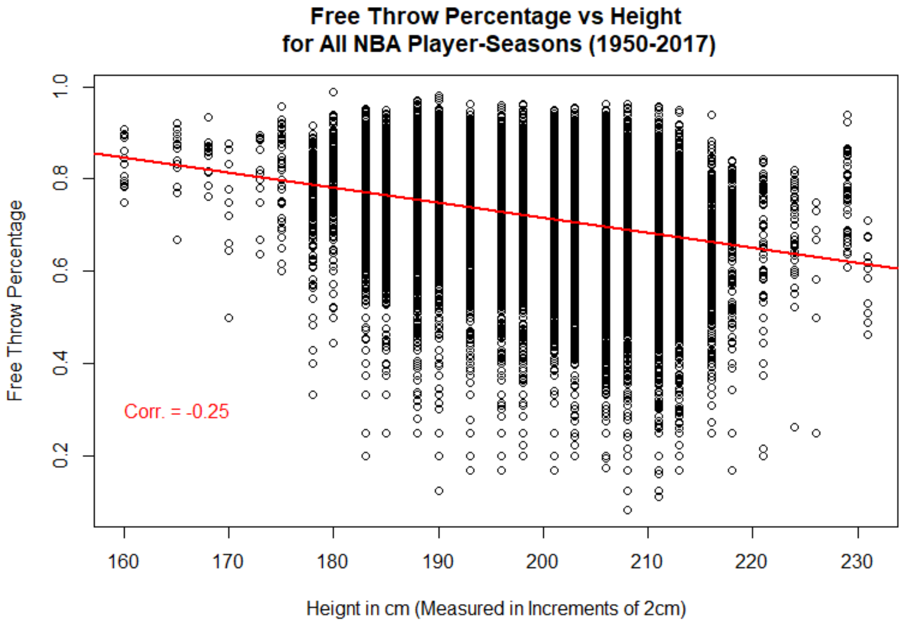
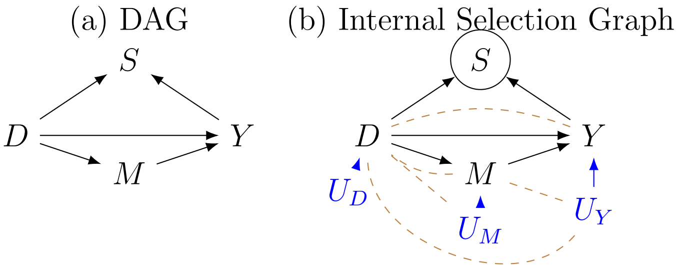


Table A.1: Free Throws and Height

Estimated Effect	SE	CI Low	CI High	DF
-0.0033	8.34e-05	-0.0034	-0.0031	22995

players typically play positions that put them closer to the basket than the free throw line. Therefore, taller players might not practice shooting from the free throw distance as much as shorter players do. Finally, there may be a form of sample selection bias sometimes called ascertainment bias. (Elwert and Winship, 2014; Schneider, 2020) It is likely that shooting ability has a positive influence on making it into the NBA. Simultaneously, it is likely that height has a positive influence on making it into the NBA. Thus, making it into the NBA is a collider between height and shooting ability. Limiting our analysis to players that made it into the NBA means we have implicitly conditioned on this collider. This can create a spurious negative relationship between height and shooting percentage.

Figure A.5: DAG and internal selection graph for height (D) and free throw percentage (Y), where M might represent a mediator like players' position and S represent an indicator for players in the NBA



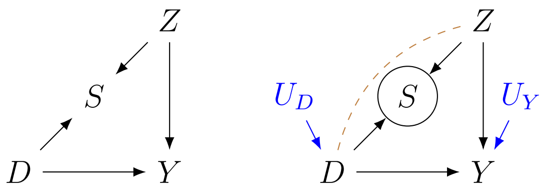
The internal selection graph for this setting might be captured by something like Figure A.5. Generalized non-causal paths between height and shooting percentage are created as a result of selecting the NBA sample. However, in this example, it actually might be reasonable to think that a negative effect of height on shooting percentage mediated by position is the driver behind the observed negative relationship. Though, as the internal selection graph makes clear, even the mediator has purely statistical relationships with the treatment and outcome due to sample selection. Paths like $D \cdots U_y \to Y$ and $D \cdots Y$ threaten the internal validity of an analysis of mediation.

A.8.2 Exercise: Zhou and Fishbach (2016)

Consider two online survey experiments from Zhou and Fishbach (2016). The authors aim to show that online experiments are often subject to high levels of attrition that can create biases. The two specific experiments we discuss here were actually *designed to induce bias resulting from attrition.* These experiments provide intuitive illustrations of how sample selection can bias treatment effect estimates, even for randomized experiments, and have the internal selection graph in Figure A.6.

Figure A.6: Zhou and Fishbach (2016) Exercise Graphs

(a) DAG (b) Internal Selection Graph



A.8.2.1 "Can doing more feel like less?"

Zhou and Fishbach (2016) "predicted that an experiment that assigns participants to recall many versus few happy events would result in a biased sample consisting of mainly happy people in the many-events condition, because happy events come to mind easily for these people, whereas the less-happy people in this condition would have to quit this difficult task. As a result of this experimental attrition, recalling many happy events could feel easier than recalling fewer happy events." The authors conducted an experiment of this form using Amazon Mechanical Turk. D is the randomly assigned treatment which consisted of being assigned to recall either 4 or 12 happy events from the last year. Y is a measure of how difficult the treatment task was to complete on a 7 point scale (1 = not difficult; 7 = extremely difficult). Z is a latent variable that captures each participant's ambient happiness. S is attrition from the study. The authors describe the attrition statistics as follows: "A total of 196 MTurk workers consented to take part in this experiment... Ninety-four of these participants dropped out of the survey once they learned what their first task (i.e., the experimental manipulation) entailed.... The dropout rate in the many condition, 69% (69/100), is significantly higher than in the few condition, 26% (25/96)..." The participants that dropped out did not complete the recollection task or rate the difficulty of the task. The mean difficulty of recall rating for the group asked to recall 12 happy events was 2.74. The mean difficulty of recall rating for the group asked to recall 4 happy events was 3.97. This is despite the fact that, "[a]ll else being equal, recalling 12 happy events from the past year requires more effort than recalling four such events." Clearly the naive treatment effect estimate is biased even for the sample of individuals that did not drop out.

The authors discuss that sample selection introduces a confound as the task is easier for happier people. This would unbalance the treatment groups. The graphical tools we have developed make this mechanism completely clear. Which treatment group you are randomly assigned to influences your decision whether to drop out or not, since one treatment is more demanding than the other and some individuals might not want to put in the extra effort. How happy you are generally influences both your decision to drop out or not, since being happier makes the task less demanding, and your rating of how difficult the task was, for the same reason. Since we are forced to condition on the selection node, a collider between the treatment assignment and happiness, a spurious relationship is created between the random treatment assignment and difficulty, which biases the estimate of the causal effect. This can easily be seen from the internal selection graph. ISAC is violated since we cannot block the path $D \cdots Z \to Y$, given that ambient happiness is unobserved.

A.8.2.2 "Can imagining applying eyeliner help one lose weight?"

Zhou and Fishbach (2016) "predicted that an experiment that assigns participants to imagine applying eyeliner (vs. applying aftershave cream) would end up with a sample that is disproportionally female. As a result, participants assigned to imagine applying eyeliner would report weighing less than those assigned to imagine applying aftershave." Obviously, there should be no direct causal effect of which question you are asked on weight. The authors again conduct an experiment of this sort using Amazon Mechanical Turk. D is the randomly assigned treatment which consisted of either being assigned to describe how applying versus not applying eyeliner would make one feel differently or being assigned to describe how applying versus not applying aftershave cream would make one feel differently. Y is the outcome and is the participants' self-reported weight in pounds. Z is an indicator variable that captures each participants gender. S is attrition from the study. The authors describe the attrition statistics as follows: “A total of 144 MTurk workers consented to take part in this experiment... Forty-one of these participants dropped out of the survey once they learned what their first task (i.e., the experimental manipulation) entailed.... The dropout rates were comparable across the two conditions: 32.4% (24/74) in the eyeliner condition and 24.3% (17/70) in the aftershave cream condition ...” The participants that dropped out did not complete the treatment task or rate the difficulty. The mean weight for the group asked to discuss eyeliner was 159 pounds. The mean weight for the group asked to discuss aftershave cream was 182 pounds. This is despite the fact that, discussing these products has no effect on weight. Clearly the naive treatment effect estimate is biased even for the sample of individuals that did not drop out. The authors observed that the eyeliner group did indeed have more females than the aftershave group.

Table A.2: Zhou and Fishbach (2016) Exercise - Average Weight by Treatment

Treatment	Average Weight	N
Eyeliner	159.02	49
Aftershave	182.08	53

The authors discuss that sample selection introduces a confound as “imagining applying eyeliner would be difficult or even aversive for average adult males, inducing them to quit.” This would unbalance the treatment groups. The graphical tools we have developed make this mechanism clear. Which treatment group you are randomly assigned to influences your decision whether to drop out or not, since one treatment is more demanding than the other depending on your gender. So your gender and the your treatment assignment both influence your decision to drop out or not. Gender also influences weight on average, given that “females generally weigh less than males.” Since we are forced to condition on the selection node, a collider between the treatment assignment and gender, a spurious relationship is created between the random treatment assignment and weight, biasing the effect estimates. This can easily be seen from the internal selection graph. ISAC is violated since we cannot block the path $D \cdots Z \rightarrow Y$, when gender is unobserved.

If gender is observed, and our causal graph is accurate, then we would be able to satisfy ISAC. Since the authors did observe participant gender, they could have adjusted for gender to get unbiased treatment effect estimates. We take this step here. Taking simple averages of weight by treatment and gender, we see that there is actually a difference in average weight between females who received the two treatments; something similar is true for males, but less so. In reality, additional variables capturing each individual's conformity to gender norms and personal preferences with respect to these products also may also be common causes of attrition and weight. So adjusting for gender may not be enough to completely identify the causal effect.

Table A.3: Zhou and Fishbach (2016) Exercise - Average Weight by Treatment and Gender

Treatment	Gender	Average Weight	N
Eyeliner	Male	182.34	29
Aftershave	Male	191.76	37
Eyeliner	Female	125.20	20
Aftershave	Female	159.69	16

We might wonder whether some other variables that might predict weight would help reduce some of the variability in weights that could be leading differences in weight, whether these differences are due purely to sampling error or to some systematic causal relationship driving the differences, like gender does. We consider adjusting for (in a simple linear model) individuals' birth year, English as first language, race, education, income, total duration of survey, and location information. All of these are pre-treatment variables and we do not believe adjusting for any of these variables will induce or amplify bias. It is possible that some are also related to attrition, like gender is. We believe adjusting for these will improve precision and reduce bias.[5] We fit a linear model with these covariates that does not adjust

[5] We assert this and the conclusions in this paragraph to keep our demonstration simple. Additional consideration might reveal more about how adjusting for these variables alters causal effect estimates.

for gender and one that does. We find that, while the model without gender still shows a positive effect on weight for individuals who discussed aftershave. The model that adjusts for gender cannot distinguish the treatment effect from zero.

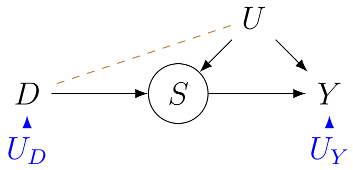
Table A.4: Zhou and Fishbach (2016) Exercise - Linear Model Treatment Effect Estimates

	Estimated Effect	SE	CI Low	CI High
Not Adjusting for Gender	20.344	9.798	0.88	39.80
Adjusting for Gender	13.937	8.705	-3.35	31.23

A.8.3 Exercise: Discrimination in Various Forms

Similar mechanisms and analysis from our working example of racial bias in policing also apply for a range of fairness and discrimination questions. The examples described by Figure A.7 are based on examples discussed in Mitchell et al. (2021). In the lending example, researchers might be attempting to understand how immigrant status relates to loan repayment or simply predicting repayment. Often the data used in such an investigation would only include individuals who actually received loans, as they have the potential to repay. In the pre-trial release example, we might want to understand how perceptions of race relate to appearance at an appointed court date or simply predicting appearance. Again, the data in such investigations are typically be limited to individuals who were actually released, as they are able to appear or not. In these examples, we again see purely statistical relationships between the protected or sensitive group statuses and outcomes that can bias the results of the investigation.

Figure A.7: Discrimination in Lending: D indicates immigrant status, S indicates receiving a loan, Y indicates repayment, and U represents unobserved factors. Discrimination in Pre-Trial Release: D indicates perception of race, S indicates pre-trial release, Y indicates appearance for court date, and U represents unobserved factors.



APPENDIX B

Appendix for Chapter 3

B.1 Traditional OVB and its reparameterization

Cinelli and Hazlett (2020) reparameterize omitted variable bias in terms of partial R^2's in the hopes of making sensitivity analysis more straight forward and the sensitivity parameters more interpretable. Traditional OVB analysis uses the Frisch–Waugh–Lovell theorem as follows.

$$\begin{aligned}
\hat{\beta}_{Y\sim D|X,S=1} &= \frac{\widehat{\text{Cov}}(D^{\perp X}, Y^{\perp X}|S=1)}{\widehat{\text{Var}}(D^{\perp X}|S=1)} \\
&= \frac{\widehat{\text{Cov}}(D^{\perp X}, \hat{\beta}_{Y\sim D|W,X,S=1}D^{\perp X} + \hat{\beta}_{Y\sim W|D,X,S=1}W^{\perp X}|S=1)}{\widehat{\text{Var}}(D^{\perp X}|S=1)} \\
&= \hat{\beta}_{Y\sim D|W,X,S=1} + \hat{\beta}_{Y\sim W|D,X,S=1}\frac{\widehat{\text{Cov}}(D^{\perp X}, W^{\perp X}|S=1)}{\widehat{\text{Var}}(D^{\perp X}|S=1)} \\
&= \hat{\beta}_{Y\sim D|W,X,S=1} + \hat{\beta}_{Y\sim W|D,X,S=1}\hat{\beta}_{W\sim D|X,S=1}
\end{aligned}$$

Cinelli and Hazlett (2020) then take the following additional steps to rewrite bias.

$$\begin{aligned}\implies \widehat{\text{bias}} &= \hat{\beta}_{Y\sim D|X,S=1} - \hat{\beta}_{Y\sim D|W,X,S=1} = \hat{\beta}_{Y\sim W|D,X,S=1}\hat{\beta}_{W\sim D|X,S=1}\\
&= \widehat{\text{Cor}}(Y^{\perp D,X}, W^{\perp D,X}|S=1)\frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\widehat{\text{SD}}(W^{\perp D,X}|S=1)}\times\\
&\widehat{\text{Cor}}(W^{\perp X}, D^{\perp X}|S=1)\frac{\widehat{\text{SD}}(W^{\perp X}|S=1)}{\widehat{\text{SD}}(D^{\perp X}|S=1)}\\
&= \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\widehat{\text{SD}}(D^{\perp X}|S=1)}\frac{\widehat{\text{SD}}(W^{\perp X}|S=1)}{\widehat{\text{SD}}(W^{\perp D,X}|S=1)}\times\\
&\widehat{\text{Cor}}(Y^{\perp D,X}, W^{\perp D,X}|S=1)\widehat{\text{Cor}}(W^{\perp X}, D^{\perp X}|S=1)\\
&= \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\widehat{\text{SD}}(D^{\perp X}|S=1)}\frac{\widehat{\text{Cor}}(Y^{\perp D,X}, W^{\perp D,X}|S=1)\widehat{\text{Cor}}(W^{\perp X}, D^{\perp X}|S=1)}{\sqrt{1-\widehat{\text{Cor}}(W^{\perp X}, D^{\perp X}|S=1)^2}}\end{aligned}$$

We can then see that the magnitude of bias can be written in terms of partial R^2's and summary information that is typical in standard OLS output.

$$\begin{aligned}\implies |\widehat{\text{bias}}| &= \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\widehat{\text{SD}}(D^{\perp X}|S=1)}\sqrt{\frac{R^2_{Y\sim W|D,X,S=1}R^2_{W\sim D|X,S=1}}{1-R^2_{W\sim D|X,S=1}}}\\
&= \text{se}(\hat{\beta}_{Y\sim D|X,S=1})\sqrt{\text{df}_{S=1}\frac{R^2_{Y\sim W|D,X,S=1}R^2_{W\sim D|X,S=1}}{1-R^2_{W\sim D|X,S=1}}}\end{aligned}$$

where $\text{se}(\hat{\beta}_{Y\sim D|X,S=1}) = \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\sqrt{\text{df}_{S=1}}\widehat{\text{SD}}(D^{\perp X}|S=1)}$ is the standard error from the regression using the selected sample and $\text{df}_{S=1}$ are that regression's degrees of freedom.

B.2 $R^2_{D\sim W|X,S=1}$ for binary random variables

Nguyen et al. (2019) provide the expression in Equation B.1 for $\text{Cov}(W, D|S=1)$ for binary variables W, D, S in their Lemma 1.

$$\text{Cov}(W, D|S = 1) = \frac{1}{P(S = 1)^2} \begin{bmatrix} P(W = 1, D = 1, S = 1)P(W = 0, D = 0, S = 1) - \\ P(W = 1, D = 0, S = 1)P(W = 0, D = 1, S = 1) \end{bmatrix} \tag{B.1}$$

To simplify things, we assume that data are generated according to the the simple collider graph: $D \rightarrow S \leftarrow W$. Nguyen et al. (2019) show that in this setting, we can write $\text{Cov}(W, D|S = 1)$ as in Equation B.2, where $P_{W=w} = P(W = w)$, $P_{D=d} = P(D = d)$, $P_{S=1} = P(S = 1)$, and $P_{S=1|wd} = P(S = 1|W = w, D = d)$.

$$\text{Cov}(W, D|S = 1) = \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{P_{S=1}^2}[P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}] \tag{B.2}$$

We can then express $\text{Cor}(W, D|S = 1)$ as follows.

$$\begin{aligned}
\text{Cor}(W, D|S = 1) &= \frac{\text{Cov}(W, D|S = 1)}{\sqrt{\text{Var}(W|S = 1)\text{Var}(D|S = 1)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]\times \\
&\quad \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{P_{S=1}^2\sqrt{\text{Var}(W|S = 1)\text{Var}(D|S = 1)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]\times \\
&\quad \sqrt{\frac{P_{W=1}^2P_{D=1}^2P_{W=0}^2P_{D=0}^2}{P_{S=1}^4P(W = 1|S = 1)P(W = 0|S = 1)P(D = 1|S = 1)P(D = 0|S = 1)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]\times \\
&\quad \sqrt{\frac{P_{W=1}^2P_{D=1}^2P_{W=0}^2P_{D=0}^2}{P(W = 1, S = 1)P(W = 0, S = 1)P(D = 1, S = 1)P(D = 0, S = 1)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]\times \\
&\quad \sqrt{\frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{P(S = 1|W = 1)P(S = 1|W = 0)P(S = 1|D = 1)P(S = 1|D = 0)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]\times \\
&\quad \sqrt{\frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{\begin{pmatrix}[P_{S=1|11}P_{D=1} + P_{S=1|10}P_{D=0}][P_{S=1|01}P_{D=1} + P_{S=1|00}P_{D=0}]\times \\ [P_{S=1|11}P_{W=1} + P_{S=1|01}P_{W=0}][P_{S=1|10}P_{W=1} + P_{S=1|00}P_{W=0}]\end{pmatrix}}}
\end{aligned}$$

So we see that $R^2_{WD|S=1}$ can be written in terms of six probabilities ($P_{S=1|11}$, $P_{S=1|00}$, $P_{S=1|10}$, $P_{S=1|01}$, $P_{W=1}$, $P_{D=1}$) as shown in Equation B.3, since $P_{W=0} = 1 - P_{W=1}$ and $P_{D=0} = 1 - P_{D=1}$.

$$\begin{aligned}
R^2_{W\sim D|S=1} &= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]^2\times \\
&\quad \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{\begin{pmatrix}[P_{S=1|11}P_{D=1} + P_{S=1|10}P_{D=0}][P_{S=1|01}P_{D=1} + P_{S=1|00}P_{D=0}]\times \\ [P_{S=1|11}P_{W=1} + P_{S=1|01}P_{W=0}][P_{S=1|10}P_{W=1} + P_{S=1|00}P_{W=0}]\end{pmatrix}}
\end{aligned} \tag{B.3}$$

The relationship between W and D in the selected sample ($S = 1$) can be expressed in terms of the relationships between S and W, D, in the full population, where we also need $P(D = 1), P(W = 1)$. In this setting, $P(S = 1)$ is actually not needed directly, since it cancelled out. All of these quantities should be easy for researchers to reason about, since they capture structural (i.e., causal) relationships between the variables.

B.3 $R^2_{D\sim W|X,S=1}$ for truncated multivariate normal random variables

To provide some intuition into how we might try to think about $R^2_{D\sim W|X,S=1}$, we consider the simple case where W, D, S have the causal structure shown in Figure 3.2 and W, D, S_0 have a multivariate normal joint distribution and $S = \mathbf{1}[S_0 \geq \mathrm{C}]$ for some $\mathrm{C} \in \mathbb{R}$. Here $X = \{\emptyset\}$. S_0 is a hypothesized latent variable that captures how W and D relate to S. The bidirected edge captures that W, D could have some relationship other than that created by conditioning on S. Within the stratum $S = 1$, we have a truncated multivariate normal joint distribution.

The post-selection covariance matrix The pre-selection covariance matrix for S_0, D, W can be written as $\Sigma = \left[\begin{array}{c|cc} \sigma^2_{S_0} & \sigma_{S_0D} & \sigma_{S_0W} \\ \hline \sigma_{S_0D} & \sigma^2_D & \sigma_{WD} \\ \sigma_{S_0W} & \sigma_{WD} & \sigma^2_W \end{array}\right] = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, where $\Sigma_{11} = \sigma^2_{S_0}$, $\Sigma_{12} = \Sigma_{21}^{\top} = \begin{bmatrix} \sigma_{S_0D} & \sigma_{S_0W} \end{bmatrix}$, and $\Sigma_{22} = \begin{bmatrix} \sigma^2_D & \sigma_{WD} \\ \sigma_{WD} & \sigma^2_W \end{bmatrix}$. Since we're interested in how the relationships between the variables change due to selection, we're interested in the covariance matrix after truncation, which can be written in terms of the pre-selection covariances:

$$\Sigma^* = \begin{bmatrix} K_{11} & K_{11}\Sigma_{11}^{-1}\Sigma_{12} \\ \Sigma_{21}\Sigma_{11}^{-1}K_{11} & \Sigma_{22} - \Sigma_{21}(\Sigma_{11}^{-1} - \Sigma_{11}^{-1}K_{11}\Sigma_{11}^{-1})\Sigma_{12} \end{bmatrix} = \begin{bmatrix} \Sigma^*_{11} & \Sigma^*_{12} \\ \Sigma^*_{21} & \Sigma^*_{22} \end{bmatrix},$$

where $K_{11} = \sigma^2_{S_0}\left[1 + \frac{\mathrm{C}\phi(\mathrm{C})}{1-\Phi(\mathrm{C})} - \left(\frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})}\right)^2\right] = \sigma^2_{S_0}\left[1 + \mathrm{C}\gamma - \gamma^2\right]$, letting $\gamma = \frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})}$, which is the inverse Mills ratio. (Kotz et al., 2000; Manjunath and Wilhelm, 2021) $S = \mathbf{1}[S_0 \geq$

C] $\iff$ $P(S=1) = P(S_0 \geq \mathrm{C}) = P(S_0 \leq -\mathrm{C}) = \Phi(-\mathrm{C})$ $\iff$ $\mathrm{C} = -\Phi^{-1}(P(S=1))$ (here we assume $S_0 \sim \mathcal{N}(0,1)$, which can be done without loss of generality; see below). $\phi(\cdot)$, $\Phi(\cdot)$, and $\Phi^{-1}(\cdot)$ are the pdf, cdf, and quantile function of the standard normal distribution. Now, we're interested in Σ^*_{22}, which contains σ^*_{DW}, the covariance between D and W after truncation.

$$\begin{aligned}
\Sigma^*_{22} &= \Sigma_{22} - \Sigma_{21}(\Sigma_{11}^{-1} - \Sigma_{11}^{-1}K_{11}\Sigma_{11}^{-1})\Sigma_{12} \\
&= \Sigma_{22} - \Sigma_{21}\left(\frac{1}{\sigma^2_{S_0}} - \frac{\sigma^2_{S_0}\left[1 + \mathrm{C}\gamma - \gamma^2\right]}{\sigma^4_{S_0}}\right)\Sigma_{12} \\
&= \Sigma_{22} - \frac{\delta}{\sigma^2_{S_0}}\Sigma_{21}\Sigma_{12}, \text{ where } \delta = \left[1 + \mathrm{C}\gamma - \gamma^2\right] \\
&= \begin{bmatrix} \sigma^2_D & \sigma_{WD} \\ \sigma_{WD} & \sigma^2_W \end{bmatrix} - \frac{\delta}{\sigma^2_{S_0}} \begin{bmatrix} \sigma_{S_0D} \\ \sigma_{S_0W} \end{bmatrix} \begin{bmatrix} \sigma_{S_0D} & \sigma_{S_0W} \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2_D & \sigma_{WD} \\ \sigma_{WD} & \sigma^2_W \end{bmatrix} - \frac{\delta}{\sigma^2_{S_0}} \begin{bmatrix} \sigma^2_{S_0D} & \sigma_{S_0D}\sigma_{S_0W} \\ \sigma_{S_0D}\sigma_{S_0W} & \sigma^2_{S_0W} \end{bmatrix} \\
\implies \sigma^*_{ab} &= \sigma_{ab} - \frac{\sigma_{S_0a}\sigma_{S_0b}}{\sigma^2_{S_0}}\delta, \ \forall a, b \in \{D, W\} \\
\implies \sigma^*_{DW} &= \sigma_{DW} - \frac{\sigma_{S_0D}\sigma_{S_0W}}{\sigma^2_{S_0}}\delta
\end{aligned}$$

We can assume $S_0 \sim \mathcal{N}(0,1)$ WOLOG Suppose that $S_0 = aD + bW + U_{S_0} = X^\top\xi + U_{S_0}$, where $U_{S_0} \sim \mathcal{N}(\mu, \sigma)$, $X = \begin{bmatrix} D & W \end{bmatrix}$, and $\xi = \begin{bmatrix} a & b \end{bmatrix}$. Since D, W, U_{S_0} are all normal random variables, so is S_0. Let $S_0' = \frac{S_0 - \mathbb{E}[S_0]}{\mathrm{SD}[S_0]} = \frac{a}{\mathrm{SD}[S_0]}D + \frac{b}{\mathrm{SD}[S_0]}W + \frac{1}{\mathrm{SD}[S_0]}U_{S_0} - \frac{\mathbb{E}[S_0]}{\mathrm{SD}[S_0]} = X^\top\xi' + \frac{1}{\mathrm{SD}[S_0]}U_{S_0} - \frac{\mathbb{E}[S_0]}{\mathrm{SD}[S_0]}$, where $\xi' = \begin{bmatrix} \frac{a}{\mathrm{SD}[S_0]} & \frac{b}{\mathrm{SD}[S_0]} \end{bmatrix}$. Since we standardized S_0 to get S_0', we know that $S_0' \sim \mathcal{N}(0,1)$. We also know that S_0' is still a linear function of D, W, and U_{S_0}. It's also easy to see how this can be extended so that X and ξ include other variables and path coefficients. Finally, we can see that

$$
\begin{aligned}
S = \mathbf{1}[S_0 \geq \mathrm{C}] = \mathbf{1}\left[\frac{S_0 - \mathbb{E}[S_0]}{\mathrm{SD}[S_0]} \geq \frac{\mathrm{C} - \mathbb{E}[S_0]}{\mathrm{SD}[S_0]}\right] = \mathbf{1}[S_0' \geq \mathrm{C}'] \\
\iff P(S = 1) = P(S_0' \geq \mathrm{C}') = \Phi(-\mathrm{C}') \iff \mathrm{C}' = -\Phi^{-1}(P(S = 1))
\end{aligned}
$$

So we can adjust the path coefficients we're considering and use S_0' rather than S_0 and just think of $S_0 \sim \mathcal{N}(0,1)$. As we saw above, we can then just consider the entire relationship between S and other variables, rather than the relationships with S_0, since we can assume $S_0 \sim \mathcal{N}(0,1)$.

Expression for $R^2_{WD|S=1}$ We can derive an expression similar to the partial correlation formula for truncated correlation and hence R^2. We can see that this is almost identical to the partial correlation formula, but for the δ's. This clarifies the difference between conditioning and truncation for normal random variables.

$$
\begin{aligned}
\rho_{WD|S=1} = \rho_{WD|S_0 \geq \mathrm{C}} = \rho^*_{WD} = \frac{\sigma^*_{WD}}{\sigma^*_D \sigma^*_W} &= \frac{\sigma_{WD} - \frac{\sigma_{S_0 D}\sigma_{S_0 W}}{\sigma^2_{S_0}}\delta}{\sqrt{\sigma^2_D - \frac{\sigma^2_{S_0 D}}{\sigma^2_{S_0}}\delta}\sqrt{\sigma^2_W - \frac{\sigma^2_{S_0 W}}{\sigma^2_{S_0}}\delta}} \\
&= \frac{\rho_{WD}\sigma_D\sigma_W - \frac{\rho_{S_0 D}\sigma_{S_0}\sigma_D\rho_{S_0 W}\sigma_{S_0}\sigma_W}{\sigma^2_{S_0}}\delta}{\sqrt{\sigma^2_D - \frac{(\rho_{S_0 D}\sigma_{S_0}\sigma_D)^2}{\sigma^2_{S_0}}\delta}\sqrt{\sigma^2_W - \frac{(\rho_{S_0 W}\sigma_{S_0}\sigma_W)^2}{\sigma^2_{S_0}}\delta}} = \frac{\sigma_D\sigma_W\left(\rho_{WD} - \rho_{S_0 D}\rho_{S_0 W}\delta\right)}{\sigma_D\sigma_W\sqrt{1 - \rho^2_{S_0 D}\delta}\sqrt{1 - \rho^2_{S_0 W}\delta}} \\
&= \frac{\rho_{WD} - \rho_{S_0 D}\rho_{S_0 W}\delta}{\sqrt{1 - \rho^2_{S_0 D}\delta}\sqrt{1 - \rho^2_{S_0 W}\delta}} \\
\rho_{WD|S_0} &= \frac{\rho_{WD} - \rho_{S_0 D}\rho_{S_0 W}}{\sqrt{1 - \rho^2_{S_0 D}}\sqrt{1 - \rho^2_{S_0 W}}}
\end{aligned}
$$

We see that the relationship between W and D in the selected (truncated) sample can be expressed in terms of the relationships between S_0, W and S_0, D as well as between W and D, in the full population. We also need $P(S = 1)$, the probability of selection or the cut point C, since $\delta = f(P(S = 1))$. If $\rho_{WD} = 0$, then $R^2_{W \sim D|S=1} = \frac{R^2_{S_0 \sim D} R^2_{S_0 \sim W}\delta}{\sqrt{1 - R^2_{S_0 \sim D}\delta}\sqrt{1 - R^2_{S_0 \sim W}\delta}}$.

Expression in terms of relationships with S, not S_0 We now explore how we can express $R^2_{WD|S=1}$ in terms of the relationships between S, W and S, D, rather than between S_0, W and S_0, D. This is useful, since here S_0 is a hypothesized latent variable, not a substantive variable. We can express ρ_{S_0W} and ρ_{S_0D} in terms of ρ_{SW} and ρ_{SD}. To see this, we borrow two results from Ding and Miratrix (2015). Assume (X_1, X_2) follows a bivariate normal with mean (μ_1, μ_2) and variance $\begin{pmatrix} \sigma_1^2 & \sigma_{12} = \rho_{12}\sigma_1\sigma_2 \\ \sigma_{12} = \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. Then for $Z_1 \sim \mathcal{N}(0,1)$, $Z_2 \sim \mathcal{N}(0,1)$, and independent from X_1, X_2 we can write

$$\begin{aligned}
X_1 &= \mu_1 + \sigma_1 Z_1 \implies Z_1 = \frac{X_1 - \mu_1}{\sigma_1} \\
X_2 &= \mu_2 + \sigma_2 \left[\rho_{12} Z_1 + \sqrt{1-\rho_{12}^2} Z_2\right] \\
&= \mu_2 + \sigma_2\rho_{12} Z_1 + \sigma_2\sqrt{1-\rho_{12}^2} Z_2 \\
&= \mu_2 + \sigma_2\rho_{12}\left[\frac{X_1-\mu_1}{\sigma_1}\right] + \sigma_2\sqrt{1-\rho_{12}^2} Z_2 \\
&= \mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1}X_1 + \sigma_2\sqrt{1-\rho_{12}^2} Z_2 \\
\implies \mathbb{E}[X_2|X_1 \geq \alpha] &= \mathbb{E}[\mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1}X_1 + \sigma_2\sqrt{1-\rho_{12}^2} Z_2 | X_1 \geq \alpha] \\
&= \mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1}\mathbb{E}[X_1|X_1 \geq \alpha] \\
\mathbb{E}[X_2|X_1 < \alpha] &= \mathbb{E}[\mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1}X_1 + \sigma_2\sqrt{1-\rho_{12}^2} Z_2 | X_1 < \alpha] \\
&= \mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1}\mathbb{E}[X_1|X_1 < \alpha]
\end{aligned}$$

$$\begin{aligned}
&\implies \mathbb{E}[X_2|X_1 \geq \alpha] - \mathbb{E}[X_2|X_1 < \alpha] \\
&= \rho_{12}\frac{\sigma_2}{\sigma_1}\left(\mathbb{E}[X_1|X_1 \geq \alpha] - \mathbb{E}[X_1|X_1 < \alpha]\right) = \rho_{12}\frac{\sigma_2}{\sigma_1}\left(\frac{f_1(\alpha)}{F_1(-\alpha)} - \frac{-f_1(\alpha)}{F_1(\alpha)}\right) \\
&= \rho_{12}\frac{\sigma_2}{\sigma_1}f_1(\alpha)\left(\frac{1}{F_1(-\alpha)} + \frac{1}{F_1(\alpha)}\right) = \rho_{12}\frac{\sigma_2}{\sigma_1}f_1(\alpha)\left(\frac{F_1(\alpha)+F_1(-\alpha)}{F_1(\alpha)F_1(-\alpha)}\right) \\
&= \rho_{12}\frac{\sigma_2}{\sigma_1}\frac{f_1(\alpha)}{F_1(\alpha)F_1(-\alpha)} = \frac{\sigma_{12}}{\sigma_1^2}\frac{f_1(\alpha)}{F_1(\alpha)F_1(-\alpha)}
\end{aligned}$$

If the marginal distribution of X_1 is $\mathcal{N}(0,1)$ then this becomes $\mathbb{E}[X_2|X_1 \geq \alpha] - \mathbb{E}[X_2|X_1 < \alpha] = \sigma_{12}\frac{\phi(\alpha)}{\Phi(\alpha)\Phi(-\alpha)}$. (Ding and Miratrix, 2015) Therefore, we have

$$\mathbb{E}[W|S=1] - \mathbb{E}[W|S=0] = \mathbb{E}[W|S_0 \geq \mathrm{C}] - \mathbb{E}[W|S_0 < \mathrm{C}] = \sigma_{S_0 W}\frac{\phi(\mathrm{C})}{\Phi(\mathrm{C})\Phi(-\mathrm{C})}$$
$$\mathbb{E}[D|S=1] - \mathbb{E}[D|S=0] = \mathbb{E}[D|S_0 \geq \mathrm{C}] - \mathbb{E}[D|S_0 < \mathrm{C}] = \sigma_{S_0 D}\frac{\phi(\mathrm{C})}{\Phi(\mathrm{C})\Phi(-\mathrm{C})}$$

For random variables X, B where $B \sim \text{Bernoulli}(p)$, $\text{Cov}(X, B) = \sigma_{xb}$ can be written as $p(1-p)\left[\mathbb{E}(X|B=1) - \mathbb{E}(X|B=0)\right]$. (Ding and Miratrix, 2015) So we see that

$$\begin{aligned}
\sigma_{SD} &= P(S=1)(1-P(S=1))\left[\mathbb{E}(D|S=1) - \mathbb{E}(D|S=0)\right] \\
&= \Phi(\mathrm{C})\Phi(-\mathrm{C})\left[\mathbb{E}(D|S=1) - \mathbb{E}(D|S=0)\right] \\
&= \Phi(\mathrm{C})\Phi(-\mathrm{C})\sigma_{DL}\frac{\phi(\mathrm{C})}{\Phi(\mathrm{C})\Phi(-\mathrm{C})} = \sigma_{S_0 D}\phi(\mathrm{C}) \\
\iff \sigma_{S_0 D} &= \frac{\sigma_{SD}}{\phi(\mathrm{C})} \\
\iff \rho_{S_0 D} &= \frac{\sigma_{SD}}{\sigma_D \sigma_{S_0}\phi(\mathrm{C})}\frac{\sigma_S}{\sigma_S} = \rho_{DS}\frac{\sigma_S}{\sigma_{S_0}\phi(\mathrm{C})} \\
&= \rho_{DS}\frac{\sqrt{P(S=1)(1-P(S=1))}}{\sigma_{S_0}\phi(\mathrm{C})} = \rho_{DS}\frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\sigma_{S_0}\phi(\mathrm{C})} = \rho_{DS}\frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\phi(\mathrm{C})}.
\end{aligned}$$

The last equality uses $S_0 \sim \mathcal{N}(0,1)$ We can do the same thing for ρ_{WL}. So we have that $\rho_{S_0 D} = \rho_{SD}\xi$ and $\rho_{S_0 W} = \rho_{SW}\xi$, where $\xi = \frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\phi(\mathrm{C})}$ can be written as a function of $P(S=1)$. We can then write

$$\rho_{WD|S=1} = \frac{\rho_{WD} - \rho_{SD}\rho_{SW}\theta}{\sqrt{1-\rho_{SD}^2\theta}\sqrt{1-\rho_{SW}^2\theta}},$$

where $\theta = \xi^2\delta$ can be written as functions of $P(S=1)$ or C. First, recall that $\xi = \frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\phi(\mathrm{C})}$, $\delta = [1 + \mathrm{C}\gamma - \gamma^2]$, and $\gamma = \frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})}$. So we can write θ in terms of C as follows or in terms of $P(S=1)$ by plugging in $\mathrm{C} = -\Phi^{-1}(P(S=1))$.

$$
\begin{aligned}
\theta = \xi^2\delta &= \left(\frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\phi(\mathrm{C})}\right)^2 \left[1 + \mathrm{C}\frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})} - \left(\frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})}\right)^2\right] \\
&= \frac{\Phi(\mathrm{C})(1-\Phi(\mathrm{C}))}{\phi(\mathrm{C})^2} + \frac{\mathrm{C}\Phi(\mathrm{C})}{\phi(\mathrm{C})} - \frac{\Phi(\mathrm{C})}{1-\Phi(\mathrm{C})}
\end{aligned}
$$

If $\rho_{WD} = 0$, then $R^2_{W\sim D|S=1} = \frac{R^2_{S\sim D}R^2_{S\sim W}\xi^2\delta}{\sqrt{1-R^2_{S\sim D}\xi^2\delta}\sqrt{1-R^2_{S\sim W}\xi^2\delta}}$. We now see that the relationship between W and D in the selected (truncated) sample can be expressed in terms of the relationships between S, W and S, D as well as between W and D, in the full population, where we also need $P(S = 1)$, the probability of selection. All of these quantities should be easy for researchers to have knowledge about and to reason about, since they capture structural (i.e., causal) relationships between the variables.

B.4 "Constant selection effects"

Suppose we would like to assume something like constant treatment effects but for R^2 between D and W after sample selection (e.g., something like $R^2_{W\sim D|S=1}$ equals $R^2_{W\sim D|S=0}$) as a way of simplifying our analysis of $R^2_{W\sim D|S=1}$. What assumptions might make sense? What expression would this provide for $R^2_{W\sim D|S=1}$? First, we expand $\mathrm{Cor}(W, D|S)$ into an expression of $\mathrm{Cor}(W, D|S = 1)$ and $\mathrm{Cor}(W, D|S = 0)$. Note that this is not a convex combination. That is the coefficients on $\mathrm{Cor}(W, D|S = 1)$ and $\mathrm{Cor}(W, D|S = 0)$ do not sum to 1.

$$
\begin{aligned}
&\mathrm{Cor}(W, D|S) \\
&= \frac{\mathrm{Cov}(W, D|S)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)} \\
&= \frac{p(S=1)\mathrm{Cov}(W, D|S=1) + p(S=0)\mathrm{Cov}(W, D|S=0)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)} \\
&= \frac{p(S=1)\mathrm{Cov}(W, D|S=1)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)} + \frac{p(S=0)\mathrm{Cov}(W, D|S=0)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)} \\
&= \frac{p(S=1)\mathrm{SD}(W|S=1)\mathrm{SD}(D|S=1)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)}\mathrm{Cor}(W, D|S=1) + \\
&\frac{p(S=0)\mathrm{SD}(W|S=0)\mathrm{SD}(D|S=0)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)}\mathrm{Cor}(W, D|S=0) \\
&= \sqrt{(A)(B)}\mathrm{Cor}(W, D|S=1) + \sqrt{(1-A)(1-B)}\mathrm{Cor}(W, D|S=0)
\end{aligned}
$$

where

$$
\begin{aligned}
A &= \frac{p(S=1)\mathrm{Var}(W|S=1)}{\mathrm{Var}(W|S)} = \frac{p(S=1)\mathrm{Var}(W|S=1)}{p(S=1)\mathrm{Var}(W|S=1) + p(S=0)\mathrm{Var}(W|S=0)} \in [0, 1] \\
B &= \frac{p(S=1)\mathrm{Var}(D|S=1)}{\mathrm{Var}(D|S)} = \frac{p(S=1)\mathrm{Var}(D|S=1)}{p(S=1)\mathrm{Var}(D|S=1) + p(S=0)\mathrm{Var}(D|S=0)} \in [0, 1]
\end{aligned}
$$

If we assume that

- $\mathrm{Cor}(W, D|S=1) = \mathrm{Cor}(W, D|S=0)$; this makes $\mathrm{Cor}(W, D|S) = \left[\sqrt{(A)(B)} + \sqrt{(1-A)(1-B)}\right] \mathrm{Cor}(W, D|S=1)$
- $\mathrm{Var}(W|S=1) = \mathrm{Var}(W|S=0)$; this makes $A = p(S=1)$
- $\mathrm{Var}(D|S=1) = \mathrm{Var}(D|S=0)$; this makes $B = p(S=1)$

These three together make $\mathrm{Cor}(W, D|S) = [p(S=1) + (1 - p(S=1))]\,\mathrm{Cor}(W, D|S=1) = \mathrm{Cor}(W, D|S=1) \implies R^2_{W\sim D|S=1} = R^2_{W\sim D|S}$. We can then leverage the partial correlation formula to arrive at

$$
R^2_{W\sim D|S=1} = R^2_{W\sim D|S} = \left(\frac{R_{W\sim D} - R_{S\sim W} R_{S\sim D}}{\sqrt{1 - R^2_{S\sim W}}\sqrt{1 - R^2_{S\sim D}}} \right)^2
$$

B.5 An often uninformative bound

In this section, we consider an bound on $R^2_{WD|S=1}$ that follows an approach similar to the last section but where we do not make the assumptions from that section. From above, we have that

$$\text{Cor}(W, D|S) = \sqrt{(A)(B)}\text{Cor}(W, D|S = 1) + \sqrt{(1 - A)(1 - B)}\text{Cor}(W, D|S = 0)$$

So we see that

$$\begin{aligned}
R^2_{W\sim D|S} &= \text{Cor}^2(W, D|S) \\
&= \left[\sqrt{(A)(B)}\text{Cor}(W, D|S = 1) + \sqrt{(1 - A)(1 - B)}\text{Cor}(W, D|S = 0)\right]^2 \\
&= \underbrace{(A)(B)R^2_{W\sim D|S=1}}_{\geq 0} + \\
&\underbrace{(1 - A)(1 - B)R^2_{W\sim D|S=0}}_{\geq 0} + \\
&2\sqrt{A(1 - A)B(1 - B)}R_{W\sim D|S=1}R_{W\sim D|S=0}
\end{aligned}$$

$\implies R^2_{W\sim D|S=1}$

$\leq \min\left(\frac{1}{AB}\left[R^2_{W\sim D|S} - 2\sqrt{A(1-A)B(1-B)}R_{W\sim D|S=1}R_{W\sim D|S=0}\right], 1\right)$

We see that $R_{W\sim D|S=1}R_{W\sim D|S=0}$ is minimized when $R_{W\sim D|S=1}R_{W\sim D|S=0} = -1$.

$\leq \min\left(\frac{1}{AB}\left[R^2_{W\sim D|S} + 2\sqrt{A(1-A)B(1-B)}\right], 1\right)$

Note that $2\sqrt{A(1-A)B(1-B)}$ is maximized at $\frac{1}{2}$ when $A = B = \frac{1}{2}$.

$= \min\left(\frac{1}{AB}\left[\left(\frac{R_{\sim WD} - R_{S\sim W}R_{S\sim D}}{\sqrt{1-R^2_{S\sim W}}\sqrt{1-R^2_{S\sim D}}}\right)^2 + 2\sqrt{A(1-A)B(1-B)}\right], 1\right)$

We can show that

$$\begin{aligned}
\mathrm{Var}(W|S) &= \mathrm{Var}(W) - \frac{\mathrm{Cov}^2(W,S)}{\mathrm{Var}(S)} = \mathrm{Var}(W)\left(1 - \frac{\mathrm{Cov}^2(W,S)}{\mathrm{Var}(W)\mathrm{Var}(S)}\right)\\
&= \mathrm{Var}(W)\left(1 - \mathrm{Cor}^2(W,S)\right) = \mathrm{Var}(W)\left(1 - R^2_{SW}\right)\\
\mathrm{Var}(D|S) &= \mathrm{Var}(D) - \frac{\mathrm{Cov}^2(D,S)}{\mathrm{Var}(S)} = \mathrm{Var}(D)\left(1 - \frac{\mathrm{Cov}^2(D,S)}{\mathrm{Var}(D)\mathrm{Var}(S)}\right)\\
&= \mathrm{Var}(D)\left(1 - \mathrm{Cor}^2(D,S)\right) = \mathrm{Var}(D)\left(1 - R^2_{SD}\right)
\end{aligned}$$

This means that

$$\begin{aligned}
A &= \frac{p(S=1)\mathrm{Var}(W|S=1)}{\mathrm{Var}(W|S)} = \frac{p(S=1)}{1-R^2_{SW}}\frac{\mathrm{Var}(W|S=1)}{\mathrm{Var}(W)} = \frac{p(S=1)}{1-R^2_{SW}}\Theta_W\\
B &= \frac{p(S=1)\mathrm{Var}(D|S=1)}{\mathrm{Var}(D|S)} = \frac{p(S=1)}{1-R^2_{SD}}\frac{\mathrm{Var}(D|S=1)}{\mathrm{Var}(D)} = \frac{p(S=1)}{1-R^2_{SD}}\Theta_D
\end{aligned}$$

So we get a bound on $R^2_{W\sim D|S=1}$:

$$R^2_{W\sim D|S=1} \leq \min\left(\frac{1}{AB}\left[\frac{(R_{W\sim D} - R_{S\sim W}R_{S\sim D})^2}{(1-R^2_{S\sim W})(1-R^2_{S\sim D})} + 2\sqrt{A(1-A)B(1-B)}\right], 1\right)$$

where $A = \frac{p(S=1)}{1-R^2_{S\sim W}}\Theta_W$, $B = \frac{p(S=1)}{1-R^2_{S\sim D}}\Theta_D$, $\Theta_W = \frac{\text{Var}(W|S=1)}{\text{Var}(W)}$, and $\Theta_D = \frac{\text{Var}(D|S=1)}{\text{Var}(D)}$.

The relationship between W and D in the selected sample can be expressed in terms of the relationships between S, W and S, D as well as between W and D, in the full population, where we also need $P(S = 1)$, Θ_W, and Θ_D. There are at least two problems with this bound. First, Θ_W and Θ_D may not be easy to reason about or to have prior knowledge about. Second, the bound is very often equal to 1. In fact, the bound very often equals 1 when $P(S = 1)$ is at all far from 1. So this is not a very useful bound.

B.6 Normalized Scaled Mutual Information Bound

Connecting $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ We start by noting that $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1}$. This is easy to see since $\eta^2_{D\sim W|S=1} = R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1} = \sup_f \left[\text{Cor}^2(D, f(W)|S = 1)\right]$. (Doksum and Samarov, 1995; Chernozhukov et al., 2022) $\eta^2_{D\sim W|S=1}$ measures portion of the variation in D that can be explained by $\mathbb{E}[D|W, S = 1]$, the conditional expectation function (CEF).[1]

Correlation and mutual information for Gaussians In order to connect $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ with mutual information, we draw inspiration from the relationship between R^2 and mutual information for random variables with Gaussian distributions. For random variables, W and D, with a bivariate Gaussian joint distribution, there is an exact relationship between R^2 (i.e., squared correlation coefficient) and mutual information (MI); see Equation B.4. (Ihara, 1993; Cover and Thomas, 2006) Can we use something like this transformation to create a useful normalized version of mutual information for non-Gaussian random variables?

[1] The law of total variance tells us that $\text{Var}(D|S = 1) = \text{Var}(\mathbb{E}[D|W, S = 1]|S = 1) + \mathbb{E}[\text{Var}(D|W, S = 1)|S = 1]$. (Aronow and Miller, 2019)

$$\mathrm{MI}(W;D) = -\frac{1}{2}\log(1 - R^2_{W\sim D}) \iff R^2_{W\sim D} = 1 - \exp(-2 \times \mathrm{MI}(W;D)) \tag{B.4}$$

A variation on the L-measure We follow the approach to normalizing mutual information laid out in Lu (2011) in transforming mutual information onto the range [0,1]. This is a variation on the transformation that holds for random variables with Gaussian joint distributions we saw in Equation B.4. Many authors have considered this type of transformation of mutual information as a way to obtain something like a non-parametric correlation based on mutual information. See Linfoot (1957); Kent (1983); Joe (1989); Kojadinovic (2005); Speed (2011); Kinney and Atwal (2014); Asoodeh et al. (2015); Smith (2015); Shevlyakov and Vasilevskiy (2017); Laarne et al. (2021). Lu (2011) introduces the L-measure. We define the squared L-measure in Equation B.5.

$$L^2(X,Y) \triangleq 1 - \exp\left(-2 \times \mathrm{IF} \times \mathrm{MI}(X;Y)\right), \tag{B.5}$$

where $\mathrm{IF} = \left(\frac{1}{1-(\mathrm{MI}(X;Y)/A)}\right)$ and $A = \sup_{U,V\in\mathcal{A}_{X,Y}} \mathrm{MI}(U;V)$.[2]

IF is a mutual information "inflation factor." We need to increase mutual information so that it goes to infinity when X, Y have a strict dependence for all types of variables, not just continuous variables. (Lu, 2011) shows that

- $A = \min[H(X), H(Y)]$, when X, Y are both discrete. This implies that $\mathrm{IF} = \left(\frac{1}{1-(\mathrm{MI}(X;Y)/\min[H(X),H(Y)])}\right) \geq 1$ since $\frac{\mathrm{MI}(X;Y)}{\min[H(X),H(Y)]} \in [0,1]$. $\mathrm{MI}(X;Y) \leq \min[H(X), H(Y)]$ since $H(X), H(Y)$ are the information content of X, Y. The idea is to inflate mutual information so that

[2] Lu (2011) defines $\mathcal{A}_{X,Y}$, U, and V in the following way: For two arbitrary random variables X and Y, with alphabet $\mathcal{X}$ and $\mathcal{Y}$, respectively, let $\mathcal{A}_{X,Y}$ be the set of all bivariate random vectors (U, V) on $\mathcal{X} \times \mathcal{Y}$ with the same marginal distributions as X and Y. Let $\mathrm{MI}(U;V)$ represent the mutual information between the random variables U and V.

IF × $\mathrm{MI}(X;Y) \to +\infty$ as X, Y become more dependent. The relationship is a strict dependence when $\mathrm{MI}(X;Y) = \min[H(X), H(Y)]$. So A gives us the right level of inflation.

- $A = H(Y)$, when Y is discrete and X is continuous. This implies that $\mathrm{IF} = \left(\frac{1}{1-(\mathrm{MI}(X;Y)/H(Y))}\right) \geq 1$. Similar ideas apply here as in the last bullet.
- $A = 1$, when X, Y are both continuous which implies that $\mathrm{IF} = 1$, since $\mathrm{MI}(X;Y) = +\infty$ for continuous variables with a strict dependence. Here no inflation is necessary.

This makes the squared L-measure is a good normalization of mutual information in that it ensures that "it is defined for any pair of random variables, it is symmetric, its value lies between 0 and 1, it equals 0 if and only if the random variables are independent, it equals 1 if there is a strict dependence between the random variables, it is invariant under marginal one-to-one transformations of the random variables, and if the random variables are Gaussian distributed, it equals" their R^2. (Lu, 2011)

For our purposes, a first question is: "does something like Equation B.6 hold?" That is, when does $R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}$ equal $1 - \exp\left(-2 \times \mathrm{IF} \times \mathrm{MI}(D; \mathbb{E}[D|W, S=1]|S=1)\right)$? We know that this would hold when D and $\mathbb{E}[D|W, S=1]$ have a Gaussian joint distribution within $S = 1$. For arbitrarily distributed variables, the relationship between D and $\mathbb{E}[D|W, S=1]$ is linear. So we would expect $R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}$ and

$$1 - \exp\left(-2 \times \mathrm{IF} \times \mathrm{MI}(D; \mathbb{E}[D|W, S=1]|S=1)\right)$$

to provide similar portraits of the dependency between D and $\mathbb{E}[D|W, S=1]$.

$$\eta^2_{D\sim W|S=1} = R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1} \overset{?}{\approx} 1 - \exp\left(-2 \times \mathrm{IF} \times \mathrm{MI}(D; \mathbb{E}[D|W, S=1]|S=1)\right) \quad \text{(B.6)}$$

$$\eta^2_{D\sim W|S=1} = R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1} = 1 - \exp\left(-2\times\mathbf{\Omega}\times\text{IF}\times\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right) \tag{B.7}$$

The question becomes whether we can alter the squared L-measure for $\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)$ to exactly recover $\eta^2_{D\sim W|S=1}$. We do this by introducing an additional mutual information scaling factor $\mathbf{\Omega} \triangleq \frac{-\frac{1}{2}\log(1-R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1})}{\text{IF}\times\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)} \geq 0$. See Equation B.7. This additional scaling factor, $\mathbf{\Omega}$, removes any discrepancy between the way that $R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}$ and the squared L-measure measure dependence between $D, \mathbb{E}[D|W,S=1]$ on the scale [0,1]. Next, the data processing inequality tells us that $\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1) \leq \text{MI}(W;D|S=1)$, since $\mathbb{E}[D|W,S=1]$ is a function of W.[3] (Cover and Thomas, 2006) It is also easy to see that $L^2_{\mathbf{\Omega}}(a) \triangleq 1-\exp\left(-2\times\mathbf{\Omega}\times\text{IF}\times a\right) \in [0,1]$ is a monotonic increasing function of $a \in [0,+\infty)$,[4] which means that $L^2_{\mathbf{\Omega}}(\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)) \leq L^2_{\mathbf{\Omega}}(\text{MI}(W;D|S=1))$. Thus, we have the relationship in Equation B.8.

$$\begin{aligned} R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} &= R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1} \\ &= L^2_{\mathbf{\Omega}}(\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)) \leq L^2_{\mathbf{\Omega}}(\text{MI}(W;D|S=1)) \end{aligned} \tag{B.8}$$

What can we say about $\mathbf{\Omega}$? Including $\mathbf{\Omega}$ in $L^2_{\mathbf{\Omega}}(\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1))$ essentially cancels out $\text{IF}\times\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)$ and undoes the L-measure transformation to simply return $R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}$. This is not a problem, since our goal is simply to find a normalization of mutual information quantities that allows us to write the bound $\eta^2_{D\sim W|S=1} = R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1} \leq L^2_{\mathbf{\Omega}}(\text{MI}(W;D|S=1))$. As we discuss in the next paragraph, we will reason about quantities like $L^2_{\mathbf{\Omega}}(\text{MI})$ directly. We do not need to directly reason about or interpret either the raw mutual information quantities, IF, or $\mathbf{\Omega}$. Moreover, due to the

[3] When the relationship between W and D is highly non-linear, $\text{MI}(W;D|S=1)$ may be much larger than $\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)$.

[4] IF in $L^2_{\mathbf{\Omega}}(a)$ is based on $\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)$.

construction of $\boldsymbol{\Omega}$, it should take values less than or equal to 1; meaning we could instead use the L-measure as a bound. This is because the transformation of $R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}$ in the numerator of $\boldsymbol{\Omega}$ is the transform that turns R^2's into mutual information for Gaussian variables. So it is an approximation to the mutual information between D and $\mathbb{E}[D|W,S=1]$, but limited to their linear relationship. If the relationship between D and $\mathbb{E}[D|W,S=1]$ is fully captured by $R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}$, then $\boldsymbol{\Omega}$ should be very close to 1. If there is some other way that D and $\mathbb{E}[D|W,S=1]$ relate, then $\boldsymbol{\Omega}$ will be less than 1, since $\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)$ captures the full relationship and IF appropriately scales mutual information for arbitrary random variables. Therefore, we might choose to consider the L-measure without scaling by $\boldsymbol{\Omega}$ either as an approximation or as a bound. Simulated examples support this discussion. See Figures 3.3 and 3.4.

Normalized scaled mutual information Our approach is to scale and normalize the mutual information using $L^2_{\boldsymbol{\Omega}}(\cdot)$. Scaling mutual information plays an important role in relating $\eta^2_{D\sim W|S=1}$ and $\text{MI}(D;W|S=1)$. We will refer to any mutual information quantity scaled by $\boldsymbol{\Omega} \times \text{IF}$ as scaled mutual information (SMI). Any mutual information quantity that is both scaled and then normalized using $L^2_{\boldsymbol{\Omega}}(\cdot)$ will be referred to as normalized scaled mutual information (NSMI). NSMI values are much easier to interpret than raw mutual information values. NSMI is a useful measure of dependence between random variables in that it satisfies the properties discussed in Rényi (1959), Smith (2015), Lu (2011), and others as the properties possessed by "an appropriate measure of dependence."[5][6]

[5] Mutual information satisfies properties 1, 2, 4, and 6. Squared Pearson correlation (i.e., R^2) satisfies properties 1, 2, 3, 5, and 7.

[6] The transformation $\ell^2(\text{MI}(X;Y)) = 1 - \exp(-2 \times \text{MI}(X;Y))$ ensures that properties 2, 3, 6, and 7 are satisfied; it is the transformation that turns mutual information into an R^2 for Gaussian distributed variables. The transformation $L^2(\text{MI}(X;Y)) = 1 - \exp(-2 \times \text{IF} \times \text{MI}(X;Y))$ is the square of Lu (2011)'s L-measure, where IF is chosen to ensure that properties 1 and 5 are satisfied, while also maintaining properties 2, 3, 6, and 7. The transformation $\text{NSMI}(X;Y) \triangleq L^2_{\boldsymbol{\Omega}}(\text{MI}(X;Y)) = 1 - \exp(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(X;Y))$ is our normalized and scaled measure of mutual information, where $\boldsymbol{\Omega} \geq 0$ is also chosen to ensure that property 8 is satisfied, while also maintaining properties 1 through 7. Lu (2011) demonstrates that properties 1 through 7 hold for the L-measure. Given this, it is trivial to see that they also hold for NSMI.

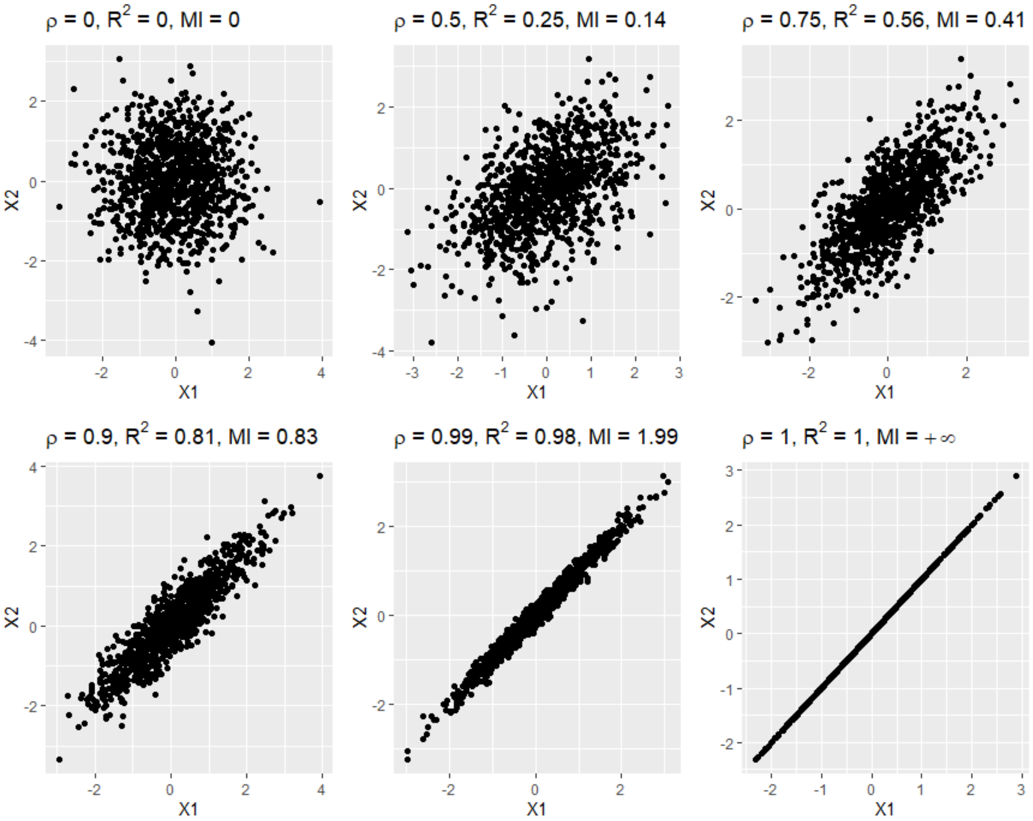
1. NSMI is defined for arbitrary pairs of random variables.
2. NSMI is symmetric.
3. NSMI takes values between 0 and 1.
4. NSMI equals 0 if and only if the variables are independent.
5. NSMI equals 1 if and only if the variables a strict dependence (functional relationship).
6. NSMI is invariant to marginal, one-to-one transformations of the variables.
7. If the variables are Gaussian distributed, then NSMI equals their R^2.[7]
8. $\text{NSMI}(D; \mathbb{E}[D|W, S=1]|S=1) = R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1} = \eta^2_{D\sim W|S=1}$.

All but the last of these are discussed in Rényi (1959), Smith (2015), and Lu (2011). The last property results from how we've defined NSMI. "Furthermore, MI is invariant under monotonic transformations of variables. This means that the MI correlation coefficient of a non-linear model (X, Y) matches the Pearson correlation of the linearized model $(f(X), g(Y))$. General conditions for f and g are described in" Ihara (1993). (Laarne et al., 2021) This statement focuses on continuous variables and the setting where the linearized model is created using monotonic transformations. $\boldsymbol{\Omega}$ will equal 1 for a linearized model. So NSMI can be interpreted as the squared Pearson correlation (i.e., R^2) of the linearized model. Figure B.2 shows the normalization curve; the normalization of SMI is precisely the normalization that turns mutual information into R^2 for Gaussian variables. Using this terminology, we see that Equation B.8 implies Equation B.9.

$$R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq \text{NSMI}(W; D|S=1) \tag{B.9}$$

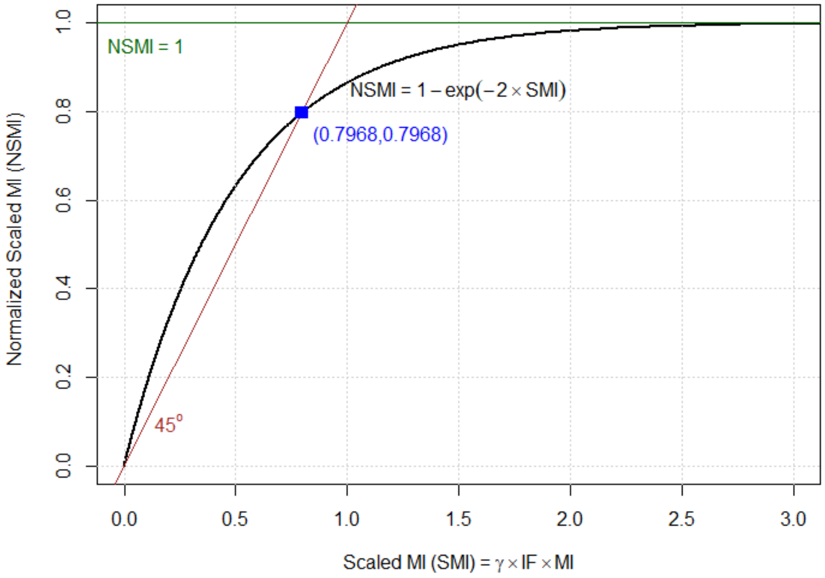
[7] It is worth noting that, although we might be more comfortable thinking about correlations and R^2's, they are not necessarily capturing what we expect. "Mutual Information is a nonlinear function of ρ which in fact makes it additive. Intuitively, in the Gaussian case, ρ should never be interpreted linearly: a ρ of $\frac{1}{2}$ carries $\approx$ 4.5 times the information of a $\rho = \frac{1}{4}$, and a ρ of $\frac{3}{4}$ 12.8 times!" (Taleb, 2019) "One needs to translate ρ into information. See how $\rho = 0.5$ is much closer to $[\rho =]0$ than to a $\rho = 1$. There are considerable differences between .9 and .99." (Taleb, 2019) See Figure B.1 for a series of plots that illustrate how changes in correlation and R^2 compare to changes in mutual information for standard Gaussian random variables. See Figure B.2 for a plot of the relationship between mutual information and R^2 for Gaussian variables, this is also the normalization curve we use.

Figure B.1: Correlation is non-linear. Scatter plots of standard Gaussian random variables with different correlations. Correlation of 0.5 is much more similar to correlation of 0 than to correlation of 1.



Interpreting NSMI For discrete random variables, X and Y, entropy and conditional entropy, $H(X)$ and $H(X|Y)$, are both positive. Recall that entropy can be thought of as a measure of the uncertainty or surprise in a random variable's outcomes. Further, $\mathrm{MI}(X;Y) = H(X) - H(X|Y)$. It is easy to see that Equation B.10 holds.

Figure B.2: Normalization of Scaled Mutual Information



$$\begin{aligned}
\mathrm{NSMI}(X;Y) &= 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(X;Y)\right) \\
&= 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times [H(X) - H(X|Y)]\right) \\
&= 1 - \exp\left([-2 \times \mathbf{\Omega} \times \mathrm{IF} \times H(X)] - [-2 \times \mathbf{\Omega} \times \mathrm{IF} \times H(X|Y)]\right) \\
&= 1 - \frac{\exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times H(X)\right)}{\exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times H(X|Y)\right)} \\
&= 1 - \frac{1 - [1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times H(X)\right)]}{1 - [1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times H(X|Y)\right)]} \\
&= 1 - \frac{1 - \mathrm{NSH}(X)}{1 - \mathrm{NSH}(X|Y)} = 1 - \frac{\mathrm{C}(X)}{\mathrm{C}(X|Y)} = \frac{\mathrm{C}(X|Y) - \mathrm{C}(X)}{\mathrm{C}(X|Y)} \\
&\text{where } \mathrm{NSH}(X) \triangleq 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times H(X)\right) \\
&\text{and } \mathrm{C}(X) \triangleq 1 - \mathrm{NSH}(X) = \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times H(X)\right)
\end{aligned} \tag{B.10}$$

NSH(X) is a normalized and scaled version of entropy that uses the same normalization and scaling as NSMI. This means that $\text{NSH}(X)$ takes values between zero and one and is a measure of the uncertainty in the outcomes of X, since it is a monotonic transformation of entropy. How might we think about the $1 - \text{NSH}(X)$ and $1 - \text{NSH}(X|Y)$ terms that appear in the expression for $\text{NSMI}(X;Y)$ in Equation B.10? $1 - \text{NSH}(X)$ close to zero means that there is a large amount of uncertainty in the outcomes of X. $1 - \text{NSH}(X)$ close to one means that there is very little uncertainty in the outcomes of X. As expected, $1 - \text{NSH}(X)$ captures something very similar to $\text{NSH}(X)$, but with the meaning of large and small values reversed. We might, therefore, call $\text{C}(X) \triangleq 1 - \text{NSH}(X)$ a measure of *lack of* surprise or *certainty*.

Thus, $\text{NSMI}(X;Y) = \frac{\text{C}(X|Y)-\text{C}(X)}{\text{C}(X|Y)}$, can be thought of as a measure of the **proportion** of the **certainty** in the outcomes of X, after we learn the value of Y, that is **gained** as a result of learning the value of Y. (As opposed to the proportion of the certainty in the outcomes of X, after we learn the value of Y, that existed before we learned the value of Y, which equals $\frac{\text{C}(X)}{\text{C}(X|Y)}$. Note that $\text{NSMI}(X;Y) + \frac{\text{C}(X)}{\text{C}(X|Y)} = 1$.)

Note that $\text{NSMI}(X;Y)$ is not a measure of the proportion of the *uncertainty* in X that is reduced by learning Y, which would be captured by $\frac{\text{NSH}(X)-\text{NSH}(X|Y)}{\text{NSH}(X)}$. But these two are closely related. Indeed, we can write $\text{NSMI}(X;Y) = \frac{\text{C}(X|Y)-\text{C}(X)}{\text{C}(X|Y)} = \frac{\text{NSH}(X)-\text{NSH}(X|Y)}{1-\text{NSH}(X|Y)}$. We see that the two share a numerator. It is only the denominator that differs. Both measure the change in information we have about X but take this as a proportion of different quantities. Note that this intuition applies to both NSMI and the L-measure.

Mutual information bounds Equation B.9 seems nice. But have we solved our original problem of finding a bound on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ in terms of structural descriptions of the relationships between the variables in the population? No we haven't. $\text{MI}(W;D|S=1)$ and $\text{NSMI}(W;D|S=1)$ both contain the spurious association between W and D created by sample selection. We now aim to find structural descriptions of the relationships between the variables in the population that can bound $\text{MI}(W;D|S=1)$. These can then be normalized to provide bounds on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$. We start by considering $\text{MI}(D;W|S)$. Using properties of mutual information (Cover and Thomas, 2006), we can show Equation B.11.

$$\begin{aligned}\mathrm{MI}(D;W|S) &= \mathrm{MI}(D;W) + \mathrm{MI}(S;D|W) - \mathrm{MI}(S;D)\\ &= \mathrm{MI}(D;W) + [\mathrm{MI}(S;[D,W]) - \mathrm{MI}(S;W)] - \mathrm{MI}(S;D)\\ &= \mathrm{MI}(D;W) + \mathrm{MI}(S;[D,W]) - \mathrm{MI}(S;D) - \mathrm{MI}(S;W)\end{aligned} \tag{B.11}$$

$\mathrm{MI}(S;[D,W]) = \mathrm{MI}(S;W) + \mathrm{MI}(S;D|W)$ is the mutual information between S and $[D,W]$ jointly. We now consider bounds on $\mathrm{MI}(D;W|S=1)$. When S is binary, two positive terms (one for $S=1$ and one for $S=0$) are being summed to create $\mathrm{MI}(D;W|S)$. See Equation B.12.

$$\begin{aligned}\mathrm{MI}(D;W|S) &= \int_{\mathcal{S}} D_{\mathrm{KL}}\left(P_{(D,W)|S} \| P_{D|S} \otimes P_{W|S}\right) dP_S\\ &= \sum_{s\in\{0,1\}} p(S=s) \sum_d \int_w p(d,w|S=s) \log\left[\frac{p(d,w|S=s)}{p(d|S=s)p(w|S=s)}\right] dd\, dw\\ &= \sum_{s\in\{0,1\}} p(S=s) D_{\mathrm{KL}}\left(P_{(D,W)|S=s} \| P_{D|S=s} \otimes P_{W|S=s}\right)\\ &= p(S=1) \times D_{\mathrm{KL}}\left(P_{(D,W)|S=1} \| P_{D|S=1} \otimes P_{W|S=1}\right)\\ &+ p(S=0) \times D_{\mathrm{KL}}\left(P_{(D,W)|S=0} \| P_{D|S=0} \otimes P_{W|S=0}\right)\\ &= p(S=1)\mathrm{MI}(D;W|S=1) + p(S=0)\mathrm{MI}(D;W|S=0)\end{aligned} \tag{B.12}$$

From Equations B.11 and B.12, we have that

$$
\begin{aligned}
\mathrm{MI}(D;W|S=1) &\leq \frac{\mathrm{MI}(D;W|S)}{p(S=1)} \\
&= \frac{\mathrm{MI}(D;W)+\mathrm{MI}(S;[D,W])-\mathrm{MI}(D;S)-\mathrm{MI}(W;S)}{p(S=1)} \\
&= \frac{\mathrm{MI}(D;W)+\mathrm{MI}(S;D|W)-\mathrm{MI}(D;S)}{p(S=1)} \\
&\leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(S;[D,W])}{p(S=1)}
\end{aligned}
\tag{B.13}
$$

This gives us the simple results in Theorem 1.

Theorem 1. *For random variables D, W, S, conditioning on S alters the relationship between D and W according to the expression $MI(D;W|S) = MI(D;W) + MI(S;[D,W]) - MI(S;D) - MI(S;W)$. Therefore, the change in dependence due to conditioning on S can be characterized using mutual information according to $MI(D;W|S) - MI(D;W) = MI(S;[D,W]) - MI(S;D) - MI(S;W)$. The dependence is not changed when $MI(S;[D,W]) = MI(S;D) + MI(S;W)$. When S is binary, it is also possible to write $MI(D;W|S) = p(S=1)MI(D;W|S=1) + p(S=0)MI(D;W|S=0)$, meaning that $MI(D;W|S=1) \leq \frac{MI(D;W|S)}{p(S=1)} = \frac{MI(D;W)+MI(S;[D,W])-MI(D;S)-MI(W;S)}{p(S=1)}$.*

So we see that we have a bound on $\mathrm{MI}(D;W|S=1)$. Every component of these bounds is something that we might have external knowledge or intuition on. From Theorem 1, we have a few relationships we can consider as bounds on $\mathrm{MI}(D;W|S=1)$. Others are also likely possible.

1. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(S;[D;W])-\mathrm{MI}(D;S)-\mathrm{MI}(W;S)}{p(S=1)}$
2. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(D;S|W)-\mathrm{MI}(D;S)}{p(S=1)}$
3. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(W;S|D)-\mathrm{MI}(W;S)}{p(S=1)}$
4. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(D;S|W)}{p(S=1)}$
5. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(W;S|D)}{p(S=1)}$

6. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(S;[D;W])}{p(S=1)}$

It is important to note that these vary in the tightness of the bound. The first three bounds are all equivalent. But the last three are not as tight, since these involve the exclusion of at least one term that is subtracted from the numerator of the first three bounds. If W and D are marginally independent, then the term $\mathrm{MI}(D;W)$ will be zero in all the bounds.

Interpretable bounds on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ We now combine Equation B.9 with Equation B.13 to get interpretable bounds on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$. We start by considering only one such bound. But others are possible.

$$\begin{aligned}
&R^2_{D\sim W|S=1}\\
&\leq \eta^2_{D\sim W|S=1} = R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}\\
&= 1-\exp\left(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right)\\
&\text{since } \boldsymbol{\Omega} = \frac{-\frac{1}{2}\log(1-R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1})}{\mathrm{IF}\times\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)}\\
&\leq 1-\exp\left(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(W;D|S=1)\right) = \mathrm{NSMI}(W;D|S=1)\\
&\text{by the data processing inequality}\\
&\leq 1-\exp\left(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\frac{\mathrm{MI}(D;W)+\mathrm{MI}(S;[D,W])}{p(S=1)}\right) \text{ by Eqn. B.13} \qquad \text{(B.14)}\\
&= 1-\exp\left(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;W)-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(S;[D,W])\right)^{\frac{1}{p(S=1)}}\\
&= 1-\left(\exp(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;W))\exp(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(S;[D,W]))\right)^{\frac{1}{p(S=1)}}\\
&= 1-([1-1+\exp(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;W))]\times\\
&[1-1+\exp(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(S;D,W]))])^{\frac{1}{p(S=1)}}\\
&= 1-([1-(1-\exp(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;W)))]\times\\
&[1-(1-\exp(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(S;[D,W])))])^{\frac{1}{p(S=1)}}\\
&= 1-\left(\ [1-\mathrm{NSMI}(D;W)][1-\mathrm{NSMI}(S;[D,W])]\ \right)^{\frac{1}{p(S=1)}}
\end{aligned}$$

Therefore, our first interpretable bound is captured by Equation B.15.

$$R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - \left(\ [1 - \mathrm{NSMI}(D;W)][1 - \mathrm{NSMI}(S;[D,W])]\ \right)^{\frac{1}{p(S=1)}} \tag{B.15}$$

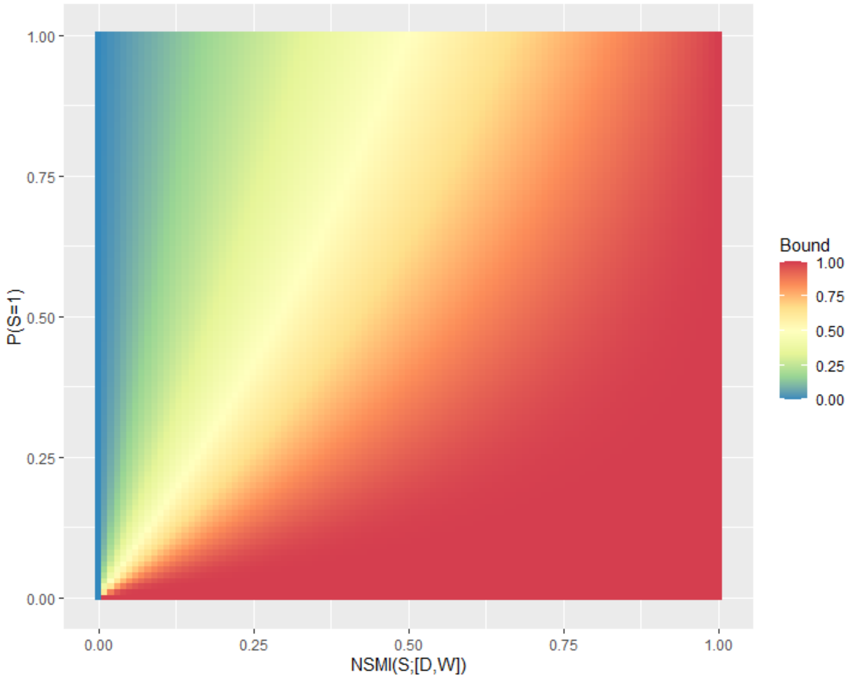
This bound is an expression of normalized scaled mutual information for the marginal mutual information between D and W, for the mutual information between S and $[D,W]$ together, and the probability of selection, $P(S = 1)$. As we saw in the case of binary random variables and truncated normal random variables, we have an expression in terms of structural (i.e., causal) relationships between the variables in the full population. In Figure B.3, we show how the bound in Equation B.15 changes for different values of $\mathrm{NSMI}(S;[D,W])$ and $p(S = 1)$. For this, we assume that that W, D are marginally independent and so $\mathrm{NSMI}(D;W) = 0$ and the bound becomes $B \overset{\Delta}{=} 1 - (1 - \mathrm{NSMI}(S;[D,W]))^{\frac{1}{p(S=1)}}$. As $p(S = 1) \to 1$, $B \to \mathrm{NSMI}(S;[D,W])$. As $p(S = 1) \to 0$, $B \to 1$. As $\mathrm{NSMI}(S;[D,W]) \to 1$, $B \to 1$. As $\mathrm{NSMI}(S;[D,W]) \to 0$, $B \to 0$. These dynamics are easy to see in the expression for the bound itself. They reflect the bounds on $\mathrm{MI}(W;D|S = 1)$ that we then scale and normalize. It is worth noting that this bound is not always informative (i.e., smaller than 1); small probabilities of selection can lead to high bounds, regardless of the value for $\mathrm{NSMI}(S;[D,W])$. This reflects that, when the selection probability is small, $\mathrm{NSMI}(S;[D,W])$ carries much less information about the stratum $S = 1$ than it does the stratum $S = 0$.

Following a similar approach as we did in obtaining the bound in Equation B.15, we arrive at Theorem 2.

Theorem 2. *For random variables D, W, and S, where S is binary, the $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ resulting after stratification to $S = 1$ can be bounded in the following ways:*

1. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(S;[D,W])]}{[1-NSMI(S;D)][1-NSMI(S;W)]}\right)^{\frac{1}{p(S=1)}}$
2. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(D;S|W)]}{[1-NSMI(S;D)]}\right)^{\frac{1}{p(S=1)}}$

Figure B.3: Bounds (from Equation B.15) on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ given values for $\text{NSMI}(S;[D,W])$ and $p(S=1)$ and assuming $\text{NSMI}(D;W)=0$



3. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(W;S|D)]}{[1-NSMI(S;W)]}\right)^{\frac{1}{p(S=1)}}$
4. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - ([1-NSMI(D;W)][1-NSMI(D;S|W)])^{\frac{1}{p(S=1)}}$
5. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - ([1-NSMI(D;W)][1-NSMI(W;S|D)])^{\frac{1}{p(S=1)}}$
6. $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1 - ([1-NSMI(D;W)][1-NSMI(S;[D,W])])^{\frac{1}{p(S=1)}}$

Bounds 1 through 3 in Theorem 2 are tighter than bounds 4 through 6, but require additional sensitivity parameters as well as some knowledge about how mutual information works. That is, since some of the NSMI quantities are related in the bounds in Theorem 2, users need to take care to reason about coherent combinations of the NSMI quantities. In particular, the bounds all take the form $1-(\tau)^{\frac{1}{p(S=1)}}$ but with different τ; τ must take a value between 0 and 1. This reflects the fact that $1-(1-\text{NSMI}(W;D|S))^{\frac{1}{p(S=1)}}$ equals bounds 1 through 3 and $\text{NSMI}(W;D|S)$ takes values between 0 and 1. This, in turn, reflects that $\text{MI}(D;W|S)=\text{MI}(D;W)+\text{MI}(S;[D,W])-\text{MI}(S;D)-\text{MI}(S;W)\geq 0$. For this reason, we encourage users unfamiliar with mutual information to use bounds 4 through 6, where the condition that $\tau \in [0, 1]$ will always be satisfied given NSMI values between 0 and 1. If W and D are assumed to be marginally independent, then $\text{NSMI}(D; W) = 0$ and this term can be removed from the bounds. Which bound is most useful depends on the relationships that practitioners feel comfortable reasoning about in terms of NSMI's.

Incorporating Covariates It is also fairly straightforward to incorporate covariates, X. We now turn to bounding $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$. The approach is very similar to the above. Equation B.16 follows from the usual expressions of $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ and the fact that $R^2_{D\sim W,X|S=1} \leq \eta^2_{D\sim W,X|S=1}$.[8]

$$\begin{aligned} R^2_{D\sim W|X,S=1} &= \frac{R^2_{D\sim W,X|S=1} - R^2_{D\sim X|S=1}}{1 - R^2_{D\sim X|S=1}} \leq \frac{\eta^2_{D\sim W,X|S=1} - R^2_{D\sim X|S=1}}{1 - R^2_{D\sim X|S=1}} \\ \eta^2_{D\sim W|X,S=1} &= \frac{\eta^2_{D\sim W,X|S=1} - \eta^2_{D\sim X|S=1}}{1 - \eta^2_{D\sim X|S=1}} \end{aligned} \tag{B.16}$$

We can estimate $R^2_{D\sim X|S=1}$ and $\eta^2_{D\sim X|S=1}$ in Equation B.16 from the selected sample, since neither involves W. Since both portions of Equation B.16 are expressions of things we can estimate from the data and $\eta^2_{D\sim W,X|S=1}$, we now turn to bounding $\eta^2_{D\sim W,X|S=1}$ in Equation B.17. Note that, as in the above discussion, $\mathbf{\Omega}$ should take values less than or equal to 1. So we could chose to omit it and simply reason about the L-measure as a bound. See the above discussion.

$$\begin{aligned} \eta^2_{D\sim W,X|S=1} &= 1 - \exp(-2 \times \mathbf{\Omega} \times \text{IF} \times \text{MI}(D; [W, X]|S = 1)) \\ \text{where } \mathbf{\Omega} &= \frac{-\frac{1}{2}\log(1 - \eta^2_{D\sim W,X|S=1})}{\text{IF} \times \text{MI}(D; [W, X]|S = 1)} \end{aligned} \tag{B.17}$$

[8] We are not able to directly link $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ as we did the versions that did not include X. If X has a very non-linear relationship with D and/or W, then it is not clear how $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ relate. In this discussion, we simply bound them separately.

From Equation B.17, we have two options for how to proceed. First, we could use Theorem 1 with W replaced with $[W, X]$ to arrive at Equation B.18.

$$\begin{aligned}\mathrm{MI}(D;[W,X]|S=1) &\leq \frac{\mathrm{MI}(D;[W,X]|S)}{p(S=1)}\\ &= \frac{\mathrm{MI}(D;[W,X]) + \mathrm{MI}(S;[D,W,X]) - \mathrm{MI}(D;S) - \mathrm{MI}([W,X];S)}{p(S=1)}\end{aligned} \tag{B.18}$$

Using Equations B.17 and B.18 we arrive at Equation B.19.

$$\begin{aligned}&\eta^2_{D\sim W,X|S=1}\\ &= 1-\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;[W,X]|S=1))\\ &\leq 1-\exp\left(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\left[\frac{\mathrm{MI}(D;[W,X]) + \mathrm{MI}(S;[D,W,X]) - \mathrm{MI}(D;S) - \mathrm{MI}([W,X];S)}{p(S=1)}\right]\right)\\ &= 1-\exp\left(-2\times\mathbf{\Omega}\times\mathrm{IF}\times[\mathrm{MI}(D;[W,X]) + \mathrm{MI}(S;[D,W,X]) - \mathrm{MI}(D;S) - \mathrm{MI}([W,X];S)]\right)^{\frac{1}{p(S=1)}}\\ &= 1-\left[\frac{\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;[W,X]))\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(S;[D,W,X]))}{\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;S))\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}([W,X];S))}\right]^{\frac{1}{p(S=1)}}\\ &= 1-\left[\frac{[1-\mathrm{NSMI}(D;[W,X])][1-\mathrm{NSMI}(S;[D,W,X])]}{[1-\mathrm{NSMI}(D;S)][1-\mathrm{NSMI}([W,X];S)]}\right]^{\frac{1}{p(S=1)}}\end{aligned} \tag{B.19}$$

Second, we could use Theorem 1 with everything conditioned on X and the fact that $\mathrm{MI}(D;W|X,S) = p(S=1)\mathrm{MI}(D;W|X,S=1) + p(S=0)\mathrm{MI}(D;W|X,S=0)$ to arrive at the second equation in Equation B.20. The first equation in Equation B.20 just comes from the definition of $\mathrm{MI}(D;[W,X]|S=1)$.

$$\begin{aligned}\mathrm{MI}(D;[W,X]|S=1) &= \mathrm{MI}(D;X|S=1) + \mathrm{MI}(D;W|X,S=1) \text{ and}\\ \mathrm{MI}(D;W|X,S=1) &\leq \frac{\mathrm{MI}(D;W|X,S)}{p(S=1)}\\ &= \frac{\mathrm{MI}(D;W|X) + \mathrm{MI}(S;[D,W]|X) - \mathrm{MI}(D;S|X) - \mathrm{MI}(W;S|X)}{p(S=1)}\end{aligned} \tag{B.20}$$

Using Equations B.17 and B.20 we arrive at Equation B.21.

$$
\begin{aligned}
&\eta^2_{D\sim W,X|S=1}\\
&= 1 - \exp(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(D;[W,X]|S=1))\\
&= 1 - \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times [\text{MI}(D;X|S=1) + \text{MI}(D;W|X,S=1)]\right)\\
&\leq 1 - \exp(-2 \times \boldsymbol{\Omega} \times \text{IF}\times\\
&\left[\text{MI}(D;X|S=1) + \frac{\text{MI}(D;W|X) + \text{MI}(S;[D,W]|X) - \text{MI}(D;S|X) - \text{MI}(W;S|X)}{p(S=1)}\right])\\
&= 1 - \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(D;X|S=1)\right)\\
&\times \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \frac{\text{MI}(D;W|X) + \text{MI}(S;[D,W]|X) - \text{MI}(D;S|X) - \text{MI}(W;S|X)}{p(S=1)}\right)\\
&= 1 - \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(D;X|S=1)\right)\\
&\times \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times [\text{MI}(D;W|X) + \text{MI}(S;[D,W]|X) - \text{MI}(D;S|X) - \text{MI}(W;S|X)]\right)^{\frac{1}{p(S=1)}}\\
&= 1 - \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(D;X|S=1)\right)\\
&\times \left[\frac{\exp(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(D;W|X))\exp(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(S;[D,W]|X))}{\exp(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(D;S|X))\exp(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(W;S|X))}\right]^{\frac{1}{p(S=1)}}\\
&= 1 - [1 - \text{NSMI}(D;X|S=1)]\left[\frac{[1-\text{NSMI}(D;W|X)][1-\text{NSMI}(S;[D,W]|X)]}{[1-\text{NSMI}(D;S|X)][1-\text{NSMI}(W;S|X)]}\right]^{\frac{1}{p(S=1)}}
\end{aligned}
\tag{B.21}
$$

Equations B.16, B.19, and B.21 combine to provide the following bounds on $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$. As before, which bound is most useful depends on what the researcher is most comfortable reasoning about.

Theorem 3. *For random variables D, W, S, X, where S is binary, the $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ resulting after stratification to $S=1$ can be bounded in the following ways, where $\lambda = \left[\frac{[1-NSMI(D;[W,X])][1-NSMI(S;[D,W,X])]}{[1-NSMI(D;S)][1-NSMI([W,X];S)]}\right]$ and $\Lambda = \left[\frac{[1-NSMI(D;W|X)][1-NSMI(S;[D,W]|X)]}{[1-NSMI(D;S|X)][1-NSMI(W;S|X)]}\right]$.*

1. $R^2_{D\sim W|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - \lambda^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right)$
2. $R^2_{D\sim W|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - [1 - NSMI(D;X|S=1)]\Lambda^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right)$
3. $\eta^2_{D\sim W|X,S=1} \leq \frac{1}{1-\eta^2_{D\sim X|S=1}} \times \left(1 - \lambda^{\frac{1}{p(S=1)}} - \eta^2_{D\sim X|S=1}\right)$

4. $\eta^2_{D\sim W|X,S=1} \leq \frac{1}{1-\eta^2_{D\sim X|S=1}} \times \left(1 - [1 - NSMI(D;X|S=1)]\Lambda^{\frac{1}{p(S=1)}} - \eta^2_{D\sim X|S=1}\right)$

where $R^2_{D\sim X|S=1}$ *or* $\eta^2_{D\sim X|S=1}$ *is estimated from the data. We can approximate or inform the choice of* $NSMI(D;X|S=1)$ *using the estimated* $R^2_{D\sim X|S=1}$, $\eta^2_{D\sim X|S=1}$, *or the related L-measure.*[9] *These bounds are all analogous to bound 1 in Theorem 2. Analogs to bounds 2 - 6 in Theorem 2 could also be formed.*

[9] We cannot directly estimate NSMI$(D;X|S=1)$, since we cannot estimate $\mathbf{\Omega}$. See Appendix B Section B.6 for discussion of $\mathbf{\Omega}$.

APPENDIX C

Appendix for Chapter 4

Here we provide technical details and prove the results found in the main text. First, we introduce a series of definitions. These are followed by a series of lemmas. Finally we state our main results in a set of theorems that follow directly from the lemmas.

C.1 Definitions

We adopt all the definitions from Appendix A, with the exception that we define internal selection graphs to focus on instruments and we define the relevance and ignorability criteria.

Definition 6 (Internal Selection Graph, G_S^+)**.** Let G be the DAG induced by a SCM.

1. Create G_S by adding an appropriately connected binary selection node, S.
2. Draw a circle around S to clearly indicate that we must limit our analysis to $S = 1$.
3. Add to G_S any node which is a parent of the treatment or a parent of a descendant of the treatment. Add to G_S any node which is a parent of the potential instrument or a parent of a descendant of the potential instrument. (U_S, the background factors contributing to selection, can be excluded.)
4. Add a dashed undirected edge between all variables between which S is a collider or an ancestor of S is a collider. We will call these dashed, undirected edges *bridges*.

Call the resulting graph an *internal selection graph*, G_S^+. (These graphs are similar to those discussed in Daniel et al. (2012) and Chapter 2.)

Definition 7 (Relevance Criterion)**.** A set of nodes X and a possible instrument IV in G_S^+

satisfy the relevance criterion relative to D (treatment), and Y (outcome) if there is at least one (*causal or generalized non-causal*) path between IV and D that does not pass through S and is not blocked by X.

Definition 8 (Ignorability Criterion)**.** A set of nodes X and a possible instrument IV in G_S^+ satisfy the ignorability criterion relative to D (treatment), and Y (outcome) if

1. No element of $\{X, S\}$ is a descendant of D and D is not in $\{X, S\}$.
2. X blocks every (*causal and generalized non-causal*) path between IV and Y except
   (a) those that pass through S and
   (b) those ending with a causal path from D to Y (e.g., paths between IV and Y that pass through D but where D or one of its descendants touches a bridge or paths on which D is an ancestor of IV must be blocked by X).

C.2 Lemmas

Lemma 25 (adapted from Shpitser et al. (2010); Pearl (1988))**.** *Let G be a causal graph. Then any model M with a distribution $P(u, v)$ inducing G, if A is d-separated from B by C in G, then A is independent of B given C, which we write $A \perp\!\!\!\perp B|C$ in $P(u, v)$.*

Lemma 26 (adapted from Shpitser et al. (2010))**.** *For every route π in G_S^+, the direct route π^* is a path. Moreover, if π is unblocked, then π^* is unblocked.*

Lemma 27. *If X and a possible instrument IV in G_S^+ satisfy the relevance criterion relative to D (treatment) and Y (outcome), then X does not d-separate D and IV in G_S^+.*

Proof. D and IV are d-separated in G_S^+ if and only if there are no unblocked paths connecting them. Consider an internal selection graph G_S^+ in which there are no paths between IV and D other than those that run through S or that are blocked by X. In such a graph, any path between IV and D that runs through S on which S is not a collider is blocked as a result of our having to condition on S and any path running through S on which S is a collider (or a descendant of a collider) corresponds to a path that is identical with the exception that the edges forming the collider are replaced with a bridge connecting the immediate parents of the collider. Any path like this that could connect D to IV must then be blocked by X by our construction. And by construction, all other paths that might connect D to IV are also blocked by X. In such a graph we can clearly see that IV and D are d-separated and we violate the relevance criterion. If we take the same graph and add one or more paths between IV and D that do not run through S and are not blocked by X, then IV and D are not d-separated and we also satisfy the relevance criterion. □

Lemma 28. *If X does not d-separate D and IV in G_S^+, then $\{X, S\}$ does not d-separate D and IV in G_S.*

Proof. If X does not d-separate D and IV in G_S^+, then there is a path that connects D to IV on which S does not appear that is not blocked by X. This is because any path that passes through S is either blocked, when S is not a collider, or, when S is a collider or a descendant of a collider, corresponds to a path that is identical except that that the edges forming the collider are replaced with a bridge connecting the immediate parents of the collider. Such paths cannot be blocked by X since X does not d-separate D and IV in G_S^+. Given a path that connects D to IV on which S does not appear and that is not blocked by X, we can find the corresponding path in G_S (which may include S as a collider or a descendant of a collider). This path will then also not be blocked when we condition on $\{X, S\}$ since S can only be either a collider or a descendant of a collider on this path in G_S and we know that X does not block it. Therefore, there is a path between D and IV in G_S that is not blocked by $\{X, S\}$ and so $\{X, S\}$ does not d-separate D and IV in G_S. □

Lemma 29. *If $\{X, S\}$ does not d-separate D and IV in G_S, then $D \not\perp\!\!\!\perp IV | X, S = 1$ for every model inducing G_S.*

Proof. This follows from Lemma 25 and the definitions of I-map and twin networks. □

Lemma 30. *If X and a possible instrument IV in G_S^+ satisfy the ignorability criterion in G_S^+ relative to D (treatment) and Y (outcome), then X d-separates IV and Y_d in N_S^+.*

Proof. We very closely follow the structure of the proof of Theorem 4 of Shpitser et al. (2010). We will show the contrapositive: assuming that we are conditioning on X, an unblocked path from IV to Y_d in N_S^+ implies that the ignorability criterion is violated in G_S^+. We will proceed in the following manner:

1. Discuss the structure of π, an unblocked path from IV to Y_d in N_S^+.
2. Discuss how sample selection relates to π.
3. Discuss a procedure for finding the path π^* in G_S^+ that corresponds to π in N_S^+.
4. Discuss possible cases for π^* in G_S^+ and their relation to the ignorability criterion.

[**1. The structure of π.**] We start by assuming that, assuming we are conditioning on X, there is an unblocked path from IV to Y_d in N_S^+. We are going to call this π. We are also able to assume, without loss of generality, that π intersects IV only at the starting point of π. What are we able to say about the structure of π across the two halves of N_S^+, the pre-intervention G_S^+ and the post-intervention $(G_S)_{\overline{D}}$? We start by noticing that the elements of Z can only appear on the pre-intervention G_S^+ side of N_S^+. This is because we can only condition on observed variables; we cannot condition on counterfactual variables, which are not observed. This means that we cannot condition on $D = d$ or any of the descendants of $D = d$ in the post-intervention $(G_S)_{\overline{D}}$ side of N_S^+. As such, as soon as π finds it way to the post-intervention side of N_S^+, the remainder of π connecting to Y_d can only contain post-intervention variables, non of which are conditioned on. Moreover, this portion of π in $(G_S)_{\overline{D}}$ can contain only edges pointing toward Y_d. This clarifies that π must be made up of first an unblocked path in the pre-intervention side, G_S^+, that we will label π_1. Next π contains one edge that points from some node in G_S^+ to some node in $(G_S)_{\overline{D}}$, which we label π_2. Recall that we are dealing with graphs in which all bidirected edges have been replaced (see the Appendix A). We will also see in the next section that π_2 cannot be a bridge. This means that the only type of edge that could connect π_1, which is entirely made up of pre-intervention nodes, to the post-intervention side is a directed edge from the pre-intervention side to the post-intervention side. An edge pointing the other direction would mean that some variables on π_1 are actually post-intervention, a contradiction. Finally, π contains the path we previously discussed, namely, a causal path that contains only descendants of $D = d$ in $(G_S)_{\overline{D}}$ that ends with Y_d. So π is composed of π_1, π_2, π_3. Since N_S^+ is built from G_S^+ and $(G_S)_{\overline{D}}$, π may contain two node "copies" that refer to the same node in G_S^+.

[**2. Sample selection and π.**] How does sample selection relate to π? Sample selection means we condition on $S = 1$. This is a pre-intervention variable. No post-intervention variable can be an ancestor of the pre-intervention version of S, otherwise we would be considering a post-intervention version of S. So all ancestors of the pre-intervention S are also pre-intervention variables. Therefore, all bridges in N_S^+ appear in the pre-intervention side of the graph, G_S^+, since we've assumed that we've replaced bidirected edges with U^*'s with uni-directional edges that point to the nodes that the bidirected edge had pointed to. Hence, any bridge on π will be in π_1. Since we must condition on the pre-intervention S, any path on which the pre-intervention S appears and is not a collider is blocked and so cannot be π. Also, any path on which the pre-intervention S appears and is a collider (or for which S is a descendant of a collider on the path) will correspond to a generalized non-causal path that is identical to the original path except that the collider is not on the generalized non-causal path and the parents of the collider are connected by a bridge on the generalized non-causal path. If the generalized non-causal path is not blocked then the original path will also not be blocked; if the generalized non-causal path is blocked then so is the original path. Therefore, we can limit our analysis to such generalized non-causal paths. So we consider π that do not contain the pre-interventional S, though π may contain bridges in π_1. Since we have assumed that sample selection is not a mediator or a descendant of a mediator, post-intervention versions of S will not appear on π_3 if they exist at all in N_S^+.

[**3. Finding the path π^* in G_S^+ that corresponds to π in N_S^+.**] How can we find a path in G_S^+ that corresponds to π? We follow a procedure laid out in Shpitser et al. (2010). First, we create π', a route in G_S^+, in this way:

1. Start by replacing each instance of a post-intervention variable in π with copy of the same node that appears on the pre-intervention side, G_S^+. We carry along the appropriate occurrence number for each of these replaced nodes.
2. Continue by replacing any instances in which the same variable appears twice in a row with only one copy of that variable. Then reduce the occurrence number of this variable by one and also do this for all the variables that follow.

The portions of π' that were created from π_1 and π_3 (portions of π in N_S^+) will also be unblocked since π_1 and π_3 are unblocked. What about the portion of π created from π_2? This will correspond to a set of three nodes where the center node is the one pointed to by π_2, which we know is a directed edge pointing to some post-intervention node, from the above discussion. The second edge in this triple must be pointing away from the middle node, since all edges in π_3 point toward Y_d, and also must be part of a causal path from D to Y in G_S^+ since the node came from the post-intervention side. But conditioning on nodes on causal paths from D to Y constitutes a violation of the ignorability criterion, so we cannot condition on the center node without violating the ignorability criterion. Therefore, this last portion of π' is also unblocked, if it exists (it may not if there are no edges in π_3). For example, say that G_S^+ contains $D \to Y$ and $D \to Z \to Y$ and sample selection is not connect to any other node. Then suppose that π is taken to be $D \to X \leftarrow U_X \to X_d \to Y_d$. Here π_1 is $D \to X \leftarrow U_X$, π_2 is the edge between U_X and X_d, and π_3 is $X_d \to Y_d$. So π' is $D \to X \leftarrow U_X \to X \to Y$. The node triple in π' that does not correspond to π_1 or π_3 is $U_X \to X \to Y$. This is blocked since we condition on X. However, conditioning on X is a violation of the ignorability criterion since X lies on a causal path from D to Y in G_S^+. All blocked versions of the node triple in π' that does not correspond to π_1 or π_3 must also violate the ignorability criterion for similar reasons. Since the middle node in this node triple is pointed to by π_2, the middle node must be a post-intervention node and so it must lie on a causal path from D to Y in G_S^+, and conditioning on it violates the ignorability criterion. Either the node triple is unblocked or it isn't. But, if it isn't, then it could only have resulted from a violation of the ignorability criterion. So π' is an unblocked route. By Lemma 26, π^*, the direct route of π' in G_S^+, is an unblocked path in G_S^+.

[4. Possible cases for π^* in G_S^+ and their relation to the ignorability criterion.]

So what are the types of π^* we might see and how do these relate to the ignorability criterion?

- If π^* does not end with a causal path from D to Y, then we immediately violate the criterion since such paths must be blocked by X.
- If π^* does end with a causal path from D to Y, then we must consider how such a π^* could have arisen from π. Since no edges can point into the post-intervention copy of D, the copy of D that we see on π^* must have resulted from the pre-intervention copy of D being on π_1. If π^* ends with a causal path from D to Y, then we assume without loss of generality that it is a proper causal path from D to Y. Since π^* ends with a causal path from D to Y, the first edge in π between D and Y must be a directed edge pointing away from an element in D. As we've discussed, π_2 must have been directed in π and pointed to a post-intervention node in $(G_S)_{\overline{D}}$ from a pre-intervention node in G_S^+. The pre-intervention node could not have been a descendant of D, otherwise it would be in the $(G_S)_{\overline{D}}$ part of N_S^+.
  - If there are no node copies that are in both π_1 and π_3 (meaning π_1 has a pre-intervention copy and π_3 has a p-intervention copy of the same node), then π^* cannot contain a proper causal path from D to Y in G_S^+. The only way it could would be for the pre-intervention node to be a descendant of D, a contradiction.
  - If there are node copies that are in both π_1 and π_3, then the only way to reach the pre-intervention node from D is via a collider unblocked by our conditioning on some element of X. This would mean that the second node in π (and the second node in π^*) is an ancestor of X, which violates the criterion.

□

Lemma 31. *If X d-separates IV and Y_d in N_S^+, then $\{X, S\}$ d-separates IV and Y_d in N_S.*

Proof. This proof take the form of the proof for similar results in Appendix A. □

Lemma 32. *If $\{X, S\}$ d-separates IV and Y_d in N_S, then $Y_d \perp\!\!\!\perp IV | X, S = 1$ for every model inducing G_s.*

Proof. This follows from Lemma 25 and the definitions of I-map and twin networks. □

C.3 Theorems

Theorem 4. *If a set of nodes X and a possible instrument IV in internal selection graph G_S^+ satisfy the relevance criterion relative to D (treatment), and Y (outcome), then $D \not\perp\!\!\!\perp IV | X, S = 1$.*

Proof. Lemmas 27, 28, and 29 prove the result. □

Theorem 5. *If a set of nodes X and a possible instrument IV in internal selection graph G_S^+ satisfy the ignorability criterion relative to D (treatment), and Y (outcome), then $Y_d \perp\!\!\!\perp IV | X, S = 1$.*

Proof. Lemmas 30, 31, and 32 prove the result. □

C.4 Violations of exclusion restriction and the definition of instruments

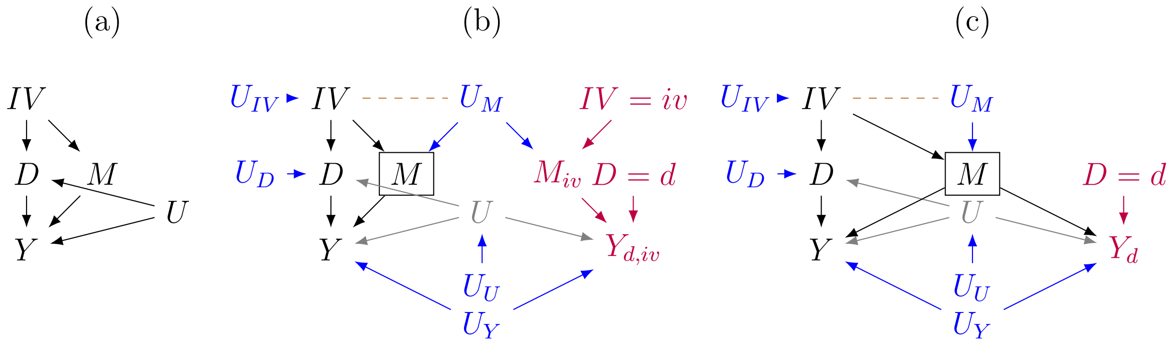
In this section we shed some light on the usefulness of the specific definition of instruments that we use. Suppose that we have the causal graph in Figure C.1(a). If we consider the definition of an instrument in which the exclusion restriction is seperated from ignorability (i.e., $Y_{d,iv} = Y_{d,iv'} = Y_d$ and $Y_{iv,d} \perp\!\!\!\perp IV$), then we have a violation of exclusion but not ignorability in this graph. (Hernán and Robins, 2006; Hernán and Robins, 2020) If we consider the definition of an instrument in which these are combined in to a single ignorability condition ($Y_d \perp\!\!\!\perp IV$ or $Y_d \perp\!\!\!\perp IV|X$), then we have a violation of ignorability.

We might consider conditioning on M to fix the problems. When we condition on M, we see that exclusion is not achieved ($Y_{d,iv} \neq Y_d$) and further we have $Y_{iv,d} \not\perp\!\!\!\perp IV|M$. However, when we condition on M, we get $Y_d \perp\!\!\!\perp IV|M$. These can be seen in Figure C.1(b,c). We're not actually interested in $Y_{d,iv}$ in its own right. There is nothing that requires that conditional ignorability statements for Y_d follow those for $Y_{d,iv}$. This is just such an example where they don't follow. Indeed, conditioning on M can actually fix problems for ignorability with Y_d but creates problems for ignorability with $Y_{d,iv}$.

So we see that while it might be intuitive to consider exclusion separately from ignorability, writing the instrument definition in this way actually imposes some unnecessary limitations on the type of conditional instruments that might work. Below, our graphical criterion allows the user to think intuitively in terms of separately ruling out causal paths and non-causal paths between IV and Y, but does not impose unnecessary restrictions as a result of writing the instrument definition in a certain way. This is a key distinction since much of the literature on instruments discusses these conditions seperately.

As we show above, we can say the following for the unconditional statements: $Y_d = Y_{iv,d}, Y_{d,iv} \perp\!\!\!\perp IV \iff Y_d \perp\!\!\!\perp IV$, but this does not hold for the conditional versions of these. In particular, $Y_d \perp\!\!\!\perp IV|M \not\Longrightarrow Y_d = Y_{iv,d}, Y_{d,iv} \perp\!\!\!\perp IV|M$. Though $Y_d = Y_{iv,d}, Y_{d,iv} \perp\!\!\!\perp IV|M \implies Y_d \perp\!\!\!\perp IV|M$.

Figure C.1: A Violation of the Exclusion Restriction



APPENDIX D

Appendix for Chapter 5

D.1 Double placebo with multiple unobserved confounders

Let us assume we have the causal graph in Figure 5.4(b). We now want to assume that there are multiple unobserved confounders captured by the vector of variables **Z**. We do not assume that the same linear combination of the components of **Z** will de-confound all the relationships we care about. $Z_{(\text{Y.DPX})} = \mathbf{Z}\beta_{Y\sim Z|D,P,X}$ is the linear combination that de-confounds the Y, D and the Y, P relationship. $Z_{(\text{N.DPX})} = \mathbf{Z}\beta_{N\sim Z|D,P,X}$ is the linear combination that de-confounds the N, D and the N, P relationship. Consider the following OVB expressions, where we use the relevant linear combination for each relationship.

$$\beta_{Y\sim D|P,X} - \beta_{Y\sim D|P,Z_{(\text{Y.DPX})},X} = \beta_{Y\sim Z_{(\text{Y.DPX})}|D,P,X}\beta_{Z_{(\text{Y.DPX})}\sim D|P,X} \tag{D.1a}$$

$$\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\text{Y.DPX})},X} = \beta_{Y\sim Z_{(\text{Y.DPX})}|D,P,X}\beta_{Z_{(\text{Y.DPX})}\sim P|D,X} \tag{D.1b}$$

$$\beta_{N\sim D|P,X} - \beta_{N\sim D|P,Z_{(\text{N.DPX})},X} = \beta_{N\sim Z_{(\text{N.DPX})}|D,P,X}\beta_{Z_{(\text{N.DPX})}\sim D|P,X} \tag{D.1c}$$

$$\beta_{N\sim P|D,X} - \beta_{N\sim P|D,Z_{(\text{N.DPX})},X} = \beta_{N\sim Z_{(\text{N.DPX})}|D,P,X}\beta_{Z_{(\text{N.DPX})}\sim P|D,X} \tag{D.1d}$$

We can re-write the middle two of these as follows.

$$\beta_{Y\sim Z_{(\mathrm{Y.DPX})}|D,P,X} = \frac{\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\mathrm{Y.DPX})},X}}{\beta_{Z_{(\mathrm{Y.DPX})}\sim P|D,X}} \tag{D.2a}$$

$$\beta_{Z_{(\mathrm{N.DPX})}\sim D|P,X} = \frac{\beta_{N\sim D|P,X} - \beta_{N\sim D|P,Z_{(\mathrm{N.DPX})},X}}{\beta_{N\sim Z_{(\mathrm{N.DPX})}|D,P,X}} \tag{D.2b}$$

Now we note that there are m_{A}, and m_{B} that satisfy the following.

$$\beta_{Z_{(\mathrm{Y.DPX})}\sim D|P,X} = m_{\mathrm{A}} \times \beta_{Z_{(\mathrm{N.DPX})}\sim D|P,X} \tag{D.3a}$$

$$\beta_{Z_{(\mathrm{Y.DPX})}\sim P|D,X} = m_{\mathrm{B}} \times \beta_{Z_{(\mathrm{N.DPX})}\sim P|D,X} \tag{D.3b}$$

Then we can get the following expression for $\beta_{Y\sim D|P,Z_{(\mathrm{Y.DPX})},X}$.

$$\begin{aligned}
&\beta_{Y\sim D|P,X} - \beta_{Y\sim D|P,Z_{(\mathrm{Y.DPX})},X} \\
&= \beta_{Y\sim Z_{(\mathrm{Y.DPX})}|D,P,X}\beta_{Z_{(\mathrm{Y.DPX})}\sim D|P,X} \\
&= m_{\mathrm{A}} \times \beta_{Y\sim Z_{(\mathrm{Y.DPX})}|D,P,X}\beta_{Z_{(\mathrm{N.DPX})}\sim D|P,X} \\
&= m_{\mathrm{A}} \times \left[\frac{\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\mathrm{Y.DPX})},X}}{\beta_{Z_{(\mathrm{Y.DPX})}\sim P|D,X}}\right]\left[\frac{\beta_{N\sim D|P,X} - \beta_{N\sim D|P,Z_{(\mathrm{N.DPX})},X}}{\beta_{N\sim Z_{(\mathrm{N.DPX})}|D,P,X}}\right] \\
&= \frac{m_{\mathrm{A}}}{m_{\mathrm{B}}} \times \frac{(\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\mathrm{Y.DPX})},X})(\beta_{N\sim D|P,X} - \beta_{N\sim D|P,Z_{(\mathrm{N.DPX})},X})}{[\beta_{N\sim P|D,X} - \beta_{N\sim P|D,Z_{(\mathrm{N.DPX})},X}]} \\
&\therefore \beta_{Y\sim D|P,Z_{(\mathrm{Y.DPX})},X} \\
&= \beta_{Y\sim D|P,X} - \frac{m_{\mathrm{A}}}{m_{\mathrm{B}}} \times \frac{(\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\mathrm{Y.DPX})},X})(\beta_{N\sim D|P,X} - \beta_{N\sim D|P,Z_{(\mathrm{N.DPX})},X})}{(\beta_{N\sim P|D,X} - \beta_{N\sim P|D,Z_{(\mathrm{N.DPX})},X})}
\end{aligned} \tag{D.4}$$

$$
\begin{aligned}
\text{Now, } \frac{m_{\mathrm{A}}}{m_{\mathrm{B}}} &= \frac{\beta_{Z_{(\mathrm{Y.DPX})}\sim D|P,X}}{\beta_{Z_{(\mathrm{N.DPX})}\sim D|P,X}} \frac{\beta_{Z_{(\mathrm{Y.DPX})}\sim P|D,X}}{\beta_{Z_{(\mathrm{N.DPX})}\sim P|D,X}} \\
&= \frac{R_{Z_{(\mathrm{Y.DPX})}\sim D|P,X} R_{Z_{(\mathrm{N.DPX})}\sim P|D,X}}{R_{Z_{(\mathrm{N.DPX})}\sim D|P,X} R_{Z_{(\mathrm{Y.DPX})}\sim P|D,X}} \times \frac{\mathrm{SD}(Z^{\perp P,X}_{(\mathrm{Y.DPX})})}{\mathrm{SD}(Z^{\perp P,X}_{(\mathrm{N.DPX})})} \frac{\mathrm{SD}(Z^{\perp D,X}_{(\mathrm{N.DPX})})}{\mathrm{SD}(Z^{\perp D,X}_{(\mathrm{Y.DPX})})} \\
&= \frac{R_{Z_{(\mathrm{Y.DPX})}\sim D|P,X} R_{Z_{(\mathrm{N.DPX})}\sim P|D,X}}{R_{Z_{(\mathrm{N.DPX})}\sim D|P,X} R_{Z_{(\mathrm{Y.DPX})}\sim P|D,X}} \times \\
&\quad \frac{\mathrm{SD}(Z^{\perp P,X}_{(\mathrm{Y.DPX})})}{\mathrm{SD}(Z^{\perp D,P,X}_{(\mathrm{Y.DPX})})} \frac{\mathrm{SD}(Z^{\perp D,P,X}_{(\mathrm{N.DPX})})}{\mathrm{SD}(Z^{\perp P,X}_{(\mathrm{N.DPX})})} \frac{\mathrm{SD}(Z^{\perp D,X}_{(\mathrm{N.DPX})})}{\mathrm{SD}(Z^{\perp D,P,X}_{(\mathrm{N.DPX})})} \frac{\mathrm{SD}(Z^{\perp D,P,X}_{(\mathrm{Y.DPX})})}{\mathrm{SD}(Z^{\perp D,X}_{(\mathrm{Y.DPX})})} \\
&= \frac{R_{Z_{(\mathrm{Y.DPX})}\sim D|P,X} R_{Z_{(\mathrm{N.DPX})}\sim P|D,X}}{R_{Z_{(\mathrm{N.DPX})}\sim D|P,X} R_{Z_{(\mathrm{Y.DPX})}\sim P|D,X}} \times \frac{\sqrt{1-R^2_{Z_{(\mathrm{N.DPX})}\sim D|P,X}}\sqrt{1-R^2_{Z_{(\mathrm{Y.DPX})}\sim P|D,X}}}{\sqrt{1-R^2_{Z_{(\mathrm{Y.DPX})}\sim D|P,X}}\sqrt{1-R^2_{Z_{(\mathrm{N.DPX})}\sim P|D,X}}} \\
&= \frac{f_{Z_{(\mathrm{Y.DPX})}\sim D|P,X} f_{Z_{(\mathrm{N.DPX})}\sim P|D,X}}{f_{Z_{(\mathrm{N.DPX})}\sim D|P,X} f_{Z_{(\mathrm{Y.DPX})}\sim P|D,X}} \\
\therefore \frac{m_{\mathrm{A}}}{m_{\mathrm{B}}} &= \frac{R_{Y\sim Z_{(\mathrm{Y.DPX})}|D,P,X} f_{Z_{(\mathrm{Y.DPX})}\sim D|P,X}}{R_{Y\sim Z_{(\mathrm{Y.DPX})}|D,P,X} f_{Z_{(\mathrm{Y.DPX})}\sim P|D,X}} \times \frac{R_{N\sim Z_{(\mathrm{N.DPX})}|D,P,X} f_{Z_{(\mathrm{N.DPX})}\sim P|D,X}}{R_{N\sim Z_{(\mathrm{N.DPX})}|D,P,X} f_{Z_{(\mathrm{N.DPX})}\sim D|P,X}} \\
&\triangleq k_{(\mathrm{YD\ /\ YP})} \times k_{(\mathrm{NP\ /\ ND})}
\end{aligned}
\tag{D.5}
$$

Thus, finally, we arrive at the expression for $\beta_{Y\sim D|P,Z_{(\mathrm{Y.DPX})},X}$ below.

$$
\begin{aligned}
\beta_{Y\sim D|P,Z_{(\mathrm{Y.DPX})},X} = \beta_{Y\sim D|P,X} - k_{(\mathrm{YD\ /\ YP})} \times k_{(\mathrm{NP\ /\ ND})} \times \\
\frac{(\beta_{Y\sim P|D,X} - \beta_{Y\sim P|D,Z_{(\mathrm{Y.DPX})},X})(\beta_{N\sim D|P,X} - \beta_{N\sim D|P,Z_{(\mathrm{N.DPX})},X})}{(\beta_{N\sim P|D,X} - \beta_{N\sim P|D,Z_{(\mathrm{N.DPX})},X})}
\end{aligned}
\tag{D.6}
$$

Some observations:

- When $Z_{(\mathrm{Y.DPX})} = Z_{(\mathrm{N.DPX})}$ or both $R_{Z_{(\mathrm{Y.DPX})}\sim D|P,X} = R_{Z_{(\mathrm{N.DPX})}\sim D|P,X}$ and $R_{Z_{(\mathrm{Y.DPX})}\sim P|D,X} = R_{Z_{(\mathrm{N.DPX})}\sim P|D,X}$, then $k_{(\mathrm{YD\ /\ YP})} \times k_{(\mathrm{NP\ /\ ND})} = 1$.
- $k_{(\mathrm{YD\ /\ YP})}$ captures the scaled ratio of the level of confounding of the Y, D relationship (conditional on P) to the level of confounding of the Y, P relationship (conditional on D), after re-scaling one of these biases, or the ratio of bias factors.
- $k_{(\mathrm{NP\ /\ ND})}$ captures the scaled ratio of the level of confounding of the N, P relationship

(conditional on D) to the level of confounding of the N, D relationship (conditional on P), after re-scaling one of these biases, or the ratio of bias factors.

- $\beta_{Y\sim P|D,Z_{(\text{Y.DPX})},X}$ measures the causal of P on Y (conditional on D).
- $\beta_{N\sim D|P,Z_{(\text{N.DPX})},X}$ measures the causal of D on N (conditional on P).
- $\beta_{N\sim P|D,Z_{(\text{N.DPX})},X}$ measures the causal of P on N (conditional on D).
- We could also use $k_{(\text{YD / ND})} \times k_{(\text{NP / YP})} =$

$$\frac{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X}f_{Z_{(\text{Y.DPX})}\sim D|P,X}}{R_{N\sim Z_{(\text{N.DPX})}|D,P,X}f_{Z_{(\text{N.DPX})}\sim D|P,X}} \times \frac{R_{N\sim Z_{(\text{N.DPX})}|D,P,X}f_{Z_{(\text{N.DPX})}\sim P|D,X}}{R_{Y\sim Z_{(\text{Y.DPX})}|D,P,X}f_{Z_{(\text{Y.DPX})}\sim P|D,X}}$$

  and have similar interpretation.

D.2 Partially linear and non-parametric framework

If investigators are interested in partially linear of fully non-parametric approaches to estimating treatment effects, we can use the omitted variables frameworks from Chernozhukov et al. (2022) to arrive at partial identification frameworks similar to those presented in the main text for settings in which placebo outcomes are available. We do not demonstrate the partially linear or non-parametric case for placebo treatments, but in principle this is also possible.

D.2.1 Partially linear framework

An investigator may be interested in estimating a partially linear model. Y is the outcome; N is a placebo outcome; D is the treatment; $\mathbf{X}$ is a vector of observed covariates; $\mathbf{Z}$ is a vector of unobserved covariates. We consider the following short and long partially linear models.

$$Y = \theta_{l,Y} D + f_l(\mathbf{X}, \mathbf{Z}) + \epsilon_l \tag{D.7a}$$

$$Y = \theta_{s,Y} D + f_s(\mathbf{X}) + \epsilon_s \tag{D.7b}$$

$$N = \theta_{l,N} D + g_l(\mathbf{X}, \mathbf{Z}) + \xi_l \tag{D.8a}$$

$$N = \theta_{s,N} D + g_s(\mathbf{X}) + \xi_s \tag{D.8b}$$

There is some m such that $\text{bias}_{(\text{YD.X})} = m \times \text{bias}_{(\text{ND.X})}$, where $\text{bias}_{(\text{YD.X})} \triangleq \theta_{s,Y} - \theta_{l,Y}$ and $\text{bias}_{(\text{ND.X})} \triangleq \theta_{s,N} - \theta_{l,N}$. This implies that Equation D.9 holds. Using Equation D.9 for partial identification will run into the issue that $m \triangleq \frac{\text{bias}_{(\text{YD.X})}}{\text{bias}_{(\text{ND.X})}}$ may be difficult to interpret when Y and N have differing scales. Can we do better?

$$\theta_{l,Y} = \theta_{s,Y} - m \times (\theta_{s,N} - \theta_{l,N}) \tag{D.9}$$

Chernozhukov et al. (2022) show that omitted variable bias in partially linear setting can be written in the following way.

$$|\theta_{s,Y} - \theta_{l,Y}|^2 = |\text{bias}_{(\text{YD.X})}|^2 = \rho_Y^2 \times S_Y^2 \times \eta^2_{Y\sim Z|D,X} \times \frac{\eta^2_{D\sim Z|X}}{1 - \eta^2_{D\sim Z|X}} \tag{D.10a}$$

$$|\theta_{s,N} - \theta_{l,N}|^2 = |\text{bias}_{(\text{ND.X})}|^2 = \rho_N^2 \times S_N^2 \times \eta^2_{N\sim Z|D,X} \times \frac{\eta^2_{D\sim Z|X}}{1 - \eta^2_{D\sim Z|X}} \tag{D.10b}$$

$\text{BF}_{(\text{YD.X})} \triangleq \eta^2_{Y\sim Z|D,X} \frac{\eta^2_{D\sim Z|X}}{1-\eta^2_{D\sim Z|X}}$ and $\text{BF}_{(\text{ND.X})} \triangleq \eta^2_{N\sim Z|D,X} \frac{\eta^2_{D\sim Z|X}}{1-\eta^2_{D\sim Z|X}}$ are "bias factors" that are scaled versions of the maximum level of unobserved confounding from $\mathbf{Z}$. S_Y^2 and S_N^2 are identified and therefore estimable from the observed data. $\eta^2_{Y\sim Z|D,X}$ and $\eta^2_{D\sim Z|X}$ are

Pearson's correlation ratios (or the non-parametric R^2s).[1] $\eta^2_{Y\sim Z|D,X}$ is the proportion of residual variation in Y explained by Z. $\eta^2_{D\sim Z|X}$ is the proportion of residual variation in D explained by Z. ρ^2_Y and ρ^2_N are referred to as "degree of adversity" in Chernozhukov et al. (2022) and reduce the actual level of bias from the maximum possible. We can now see that we can re-express m as in Equation D.11.

$$\begin{aligned}|m|^2 = \frac{|\text{bias}_{(\text{YD.X})}|^2}{|\text{bias}_{(\text{ND.X})}|^2} &= \frac{\rho^2_Y \times S^2_Y \times \eta^2_{Y\sim Z|D,X} \times \frac{\eta^2_{D\sim Z|X}}{1-\eta^2_{D\sim Z|X}}}{\rho^2_N \times S^2_N \times \eta^2_{N\sim Z|D,X} \times \frac{\eta^2_{D\sim Z|X}}{1-\eta^2_{D\sim Z|X}}} \\ &= \frac{\rho^2_Y}{\rho^2_N} \times \frac{\eta^2_{Y\sim Z|D,X} \times \frac{\eta^2_{D\sim Z|X}}{1-\eta^2_{D\sim Z|X}}}{\eta^2_{N\sim Z|D,X} \times \frac{\eta^2_{D\sim Z|X}}{1-\eta^2_{D\sim Z|X}}} \times \frac{S^2_Y}{S^2_N} = \gamma \times k \times \frac{S^2_Y}{S^2_N}\end{aligned} \tag{D.11}$$

$\gamma \triangleq \frac{\rho^2_Y}{\rho^2_N}$ and $k \triangleq \frac{\eta^2_{Y\sim Z|D,X} \times \frac{\eta^2_{D\sim Z|X}}{1-\eta^2_{D\sim Z|X}}}{\eta^2_{N\sim Z|D,X} \times \frac{\eta^2_{D\sim Z|X}}{1-\eta^2_{D\sim Z|X}}} = \frac{\eta^2_{Y\sim Z|D,X}}{\eta^2_{N\sim Z|D,X}}$. Which means that we can write Equation D.12. We could use Equation D.12 for partial identification by reasoning about plausible ranges of values for k, γ, and $\theta_{l,N}$, while estimating the remaining terms from observed data. k represents the relative level of confounding where differences in scale have been accounted for. $\theta_{l,N}$ is the effect of D on N. γ is the relative level of adversity. k and $\theta_{l,N}$ should be easy to reason about using external knowledge. It is possible that γ would be more difficult to consider. We leave this discussion here for now, however. A reasonable simplifying assumption may be that γ is close to 1. See Chernozhukov et al. (2022) for further discussion of η^2s, S^2_Y, S^2_N, ρ^2_Y, and ρ^2_N.

$$\theta_{l,Y} = \theta_{s,Y} - \text{sign}(m) \times \sqrt{\gamma \times k} \times (\theta_{s,N} - \theta_{l,N}) \times \sqrt{\frac{S^2_Y}{S^2_N}} \tag{D.12}$$

[1] $\eta^2_{D\sim Z|X} = \frac{\text{Var}(\mathbb{E}[D|Z,X]) - \text{Var}(\mathbb{E}[D|X])}{\text{Var}(D) - \text{Var}(\mathbb{E}[D|X])} = \frac{\eta^2_{D\sim Z,X} - \eta^2_{D\sim X}}{1-\eta^2_{D\sim X}}$. $\eta^2_{Y\sim Z|D,X}$ can be similarly interpreted.

D.2.2 Non-parametric framework

An investigator may also be interested in estimating a linear functional of the conditional expectation function of the outcome in a fully non-parametric setting, like $\theta_{l,Y} = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[f_Y(1, \mathbf{X}, \mathbf{Z}) - f_Y(0, \mathbf{X}, \mathbf{Z})]$ for a binary treatment D, where $Y_d = f_Y(d, \mathbf{X}, \mathbf{Z}, U_Y)$ is the equation for Y in the structural causal model under intervention to set $D = d$. Again, the investigator is only able to estimate $\theta_{s,Y} = \mathbb{E}[f_Y^*(1, \mathbf{X}) - f_Y^*(0, \mathbf{X})]$, where $f_Y^*(D, X) \triangleq \mathbb{E}[Y|D, \mathbf{X}] = \mathbb{E}[f_Y(D, X, \mathbf{Z})|D, \mathbf{X}]$. We define $\theta_{l,N}$ and $\theta_{s,N}$ similarly for a placebo outcome. We could again consider an approach like Equation D.13. But again $m \triangleq \frac{\text{bias}_{(\text{YD.X})}}{\text{bias}_{(\text{ND.X})}}$ may be difficult to interpret when Y and N have differing scales.

$$\theta_{l,Y} = \theta_{s,Y} - m \times (\theta_{s,N} - \theta_{l,N}) \tag{D.13}$$

Chernozhukov et al. (2022) also show that omitted variable bias in fully non-parametric setting can be written as an expression in terms of $\eta^2_{Y\sim Z|D,X}$ and a second term that, in the case of targeting $\theta_{l,Y} = \mathbb{E}[Y_1 - Y_0]$ with a binary treatment D, is the "average gain in the conditional precision with which we predict D by using Z in addition to X," which is somewhat similar to $\eta^2_{D\sim Z|X}$. Specifically, we have the following.

$$|\theta_{s,Y} - \theta_{l,Y}|^2 = |\text{bias}_{(\text{YD.X})}|^2 = \rho_Y^2 \times S_Y^2 \times \eta^2_{Y\sim Z|D,X} \times \frac{1 - R^2_{\alpha\sim\alpha_s}}{R^2_{\alpha\sim\alpha_s}} \tag{D.14a}$$

$$|\theta_{s,N} - \theta_{l,N}|^2 = |\text{bias}_{(\text{YD.X})}|^2 = \rho_N^2 \times S_N^2 \times \eta^2_{N\sim Z|D,X} \times \frac{1 - R^2_{\alpha\sim\alpha_s}}{R^2_{\alpha\sim\alpha_s}} \tag{D.14b}$$

$\text{BF}_{(\text{YD.X})} \triangleq \eta^2_{Y\sim Z|D,X} \frac{1-R^2_{\alpha\sim\alpha_s}}{R^2_{\alpha\sim\alpha_s}}$ and $\text{BF}_{(\text{ND.X})} \triangleq \eta^2_{N\sim Z|D,X} \frac{1-R^2_{\alpha\sim\alpha_s}}{R^2_{\alpha\sim\alpha_s}}$ are "bias factors" that are scaled versions of the maximum level of unobserved confounding from $\mathbf{Z}$. S_Y^2 and S_N^2 are identified and therefore estimable from the data. See Chernozhukov et al. (2022) for discussion of S_Y^2, S_N^2, and $R^2_{\alpha\sim\alpha_s}$. Proceeding as in the last section, we can arrive at Equation

D.15. Here again $k = \frac{\eta^2_{Y\sim Z|D,X}}{\eta^2_{N\sim Z|D,X}}$.

$$\theta_{l,Y} = \theta_{s,Y} - \operatorname{sign}(m) \times \sqrt{\gamma \times k} \times (\theta_{s,N} - \theta_{l,N}) \times \sqrt{\frac{S_Y^2}{S_N^2}} \tag{D.15}$$

D.3 Re-expression of bias factors

$$\begin{aligned}
\beta_{Y\sim D|X} - \beta_{Y\sim D|Z,X} &= R_{Y\sim Z|D,X} f_{D\sim Z|X} \frac{\text{SD}(Y^{\perp D,X})}{\text{SD}(D^{\perp X})} \\
\implies R_{Y\sim Z|D,X} f_{D\sim Z|X} &= (\beta_{Y\sim D|X} - \beta_{Y\sim D|Z,X}) \frac{\text{SD}(D^{\perp X})}{\text{SD}(Y^{\perp D,X})} \\
&= R_{Y\sim D|X} \frac{\text{SD}(Y^{\perp X})}{\text{SD}(D^{\perp X})} \frac{\text{SD}(D^{\perp X})}{\text{SD}(Y^{\perp D,X})} \\
&- R_{Y\sim D|Z,X} \frac{\text{SD}(Y^{\perp X,Z})}{\text{SD}(D^{\perp X,Z})} \frac{\text{SD}(D^{\perp X})}{\text{SD}(Y^{\perp D,X})} \frac{\text{SD}(Y^{\perp D,X,Z})}{\text{SD}(Y^{\perp D,X,Z})} \\
&= f_{Y\sim D|X} - f_{Y\sim D|Z,X} \frac{\text{SD}(Y^{\perp D,X,Z})}{\text{SD}(D^{\perp X,Z})} \frac{\text{SD}(D^{\perp X})}{\text{SD}(Y^{\perp D,X})} \\
&= f_{Y\sim D|X} - f_{Y\sim D|Z,X} \sqrt{\frac{1 - R^2_{Y\sim Z|D,X}}{1 - R^2_{D\sim Z|X}}} \\
\implies 1 &= \frac{f_{Y\sim D|X}}{R_{Y\sim Z|D,X} f_{D\sim Z|X}} - \frac{f_{Y\sim D|Z,X}}{f_{Y\sim Z|D,X} R_{D\sim Z|X}}
\end{aligned}$$

Bibliography

Amorim, L. (2022). Replication Data for: Zika Epidemic and Birth Rates in Brazil.

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 Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arah, O. A. (2019). Analyzing selection bias for credible causal inference. *Epidemiology*, 30(4):517–520.

Arnold, B. F. and Ercumen, A. (2016). Negative Control Outcomes: A Tool to Detect Bias in Randomized Trials. *JAMA*, 316(24):2597–2598.

Arnold, B. F., Ercumen, A., Benjamin-Chung, J., and John M. Colford, J. (2016). Negative Controls to Detect Selection Bias and Measurement Bias in Epidemiologic Studies. *Epidemiology*, 27(5):637–641.

Aronow, P. M. and Miller, B. T. (2019). *Foundations of Agnostic Statistics*. Cambridge University Press.

Asoodeh, S., Alajaji, F., and Linder, T. (2015). On maximal correlation, mutual information and data privacy. In *2015 IEEE 14th Canadian Workshop on Information Theory (CWIT)*, pages 27–31.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33:2297–2340.

Balke, A. and Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth International Conference on Uncertainty in*

Artificial Intelligence, UAI'94, page 46–54, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Balke, A. and Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. In *AAAI*.

Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. In Lawrence, N. D. and Girolami, M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 100–108, La Palma, Canary Islands. PMLR.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Bareinboim, E., Tian, J., and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*, 6(51):1173–82.

Berk, R. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, 48(3):386–398.

Bilinski, A. and Hatfield, L. A. (2018). Nothing to see here? non-inferiority approaches to parallel trends and other model assumptions.

Callaway, B. (2019). *qte: Quantile Treatment Effects*. R package version 1.3.0.

Campbell, D. T. (1957). Factors relevant to validity of experiments in social settings. *Psychological Bulletin*, 54(4):297–312.

Campbell, D. T. and Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago.

Campbell, D. T. and Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago.

Canan, C., Lesko, C., and Lau, B. (2017). Instrumental Variable Analyses and Selection Bias. *Epidemiology*, 28(3):396–398.

Chabé-Ferret, S. (2017). Should We Combine Difference In Differences with Conditioning on Pre-Treatment Outcomes? TSE Working Papers 17-824, Toulouse School of Economics (TSE).

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., and Syrgkanis, V. (2022). Long story short: Omitted variable bias in causal machine learning. Working Paper 30302, National Bureau of Economic Research.

Cinelli, C., Ferwerda, J., and Hazlett, C. (2020). sensemakr: Sensitivity analysis tools for ols in r and stata.

Cinelli, C., Forney, A., and Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*, 0(0):00491241221099552.

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin.

Correa, J. and Bareinboim, E. (2017). Causal effect identification by adjustment under confounding and selection biases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Correa, J., Tian, J., and Bareinboim, E. (2018). Generalized adjustment under confounding and selection biases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Correa, J., Tian, J., and Bareinboim, E. (2019). Adjustment criteria for generalizing experimental findings. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1361–1369. PMLR.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Cuddeback, G., Wilson, E., Orme, J. G., and Combs-Orme, T. (2004). Detecting and statistically correcting sample selection bias. *Journal of Social Service Research*, 30(3):19–33.

Daniel, R. M., Kenward, M. G., Cousens, S. N., and De Stavola, B. L. (2012). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256.

Department of Labor, U. S. (2023). History of changes to the minimum wage law.

Didelez, V., Kreiner, S., and Keiding, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387.

Didelez, V. and Sheehan, N. (2007a). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330.

Didelez, V. and Sheehan, N. A. (2007b). Mendelian randomisation: Why epidemiology needs a formal language for causality. In Russo, F. and Williamson, J., editors, *Causality and Probability in the Sciences*, pages 5–263.

Ding, P. and Miratrix, L. W. (2015). To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference*, 3(1):41–57.

Doksum, K. and Samarov, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, 23(5):1443–1473.

Egami, N. and Hartman, E. (2021). Covariate selection for generalizing experimental results: Application to a large-scale development program in uganda*. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(4):1524–1548.

Egami, N. and Hartman, E. (2022). Elements of external validity: Framework, design, and analysis. *American Political Science Review*, page 1–19.

Elwert, F. and Segarra, E. (2022). *Instrumental Variables with Treatment-Induced Selection: Exact Bias Results*, page 575–592. Association for Computing Machinery, New York, NY, USA, 1 edition.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40(1):31–53. PMID: 30111904.

Ertefaie, A., Small, D., Flory, J., and Hennessy, S. (2016). Selection bias when using instrumental variable methods to compare two treatments but more than two treatments are available. *The International Journal of Biostatistics*, 12(1):219–232.

Ferwerda, J., Hainmueller, J., and Hazlett, C. J. (2017). Kernel-based regularized least squares in r (krls) and stata (krls). *Journal of Statistical Software*, 79(3):1–26.

Flanders, W. D. and Ye, D. (2019). Limits for the magnitude of m-bias and certain other types of structural selection bias. *Epidemiology*, 30(4):501–508.

Freidling, T. and Zhao, Q. (2023). Sensitivity analysis with the r^2-calculus.

Freyaldenhoven, S., Hansen, C., and Shapiro, J. M. (2019). Pre-event trends in the panel event-study design. *American Economic Review*, 109(9):3307–38.

Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3:151–182.

Ghassami, A. and Kiyavash, N. (2017). Interaction information for causal inference: The case of directed triangle. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1326–1330.

Gibson, L. and Zimmerman, F. (2021). Measuring the sensitivity of difference-in-difference estimates to the parallel trends assumption. *Research Methods in Medicine & Health Sciences*, 2(4):148–156.

Gkatzionis, A. and Burgess, S. (2018). Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *International Journal of Epidemiology*, 48(3):691–701.

Greenland, S. (1977). Response and follow-up bias in cohort studies. *American Journal of Epidemiology*, 106(3):184–187.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306.

Greenland, S. (2022). *The Causal Foundations of Applied Probability and Statistics*, page 605–624. Association for Computing Machinery, New York, NY, USA, 1 edition.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Greiner, D. J. and Rubin, D. B. (2011). Causal Effects of Perceived Immutable Characteristics. *The Review of Economics and Statistics*, 93(3):775–785.

Grogger, J. and Ridgeway, G. (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101(475):878–887.

Hainmueller, J. and Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 22(2):143–168.

Ham, D. W. and Miratrix, L. (2022). Benefits and costs of matching prior to a difference in differences analysis when parallel trends does not hold.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):757–778.

Hausman, J. A. and Wise, D. A. (1977). Social experimentation, truncated distributions, and efficient estimation. *Econometrica*, 45(4):919–938.

Hazlett, C. (2020). Angry or weary? how violence impacts attitudes toward peace among darfurian refugees. *Journal of Conflict Resolution*, 64(5):844–870.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Helin, K. (2011). Free throw shooting: It's not about where you're from, it's about height. https://nba.nbcsports.com/2011/03/31/free-throw-shooting-it%E2%80%99s-not-about-where-you%E2%80%99re-from-it%E2%80%99s-about-height/. [Online; posted 31-March-2011].

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586.

Hernán, M., Hernández-Díaz, S., and Robins, J. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

Hernán, M. and Robins, J. (2020). *Causal Inference: What If.* Chapman & Hall/CRC, Boca Raton.

Hernán, M. A. (2017). Invited Commentary: Selection Bias Without Colliders. *American Journal of Epidemiology*, 185(11):1048–1050.

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

Howe, C. J., Cain, L. E., and Hogan, J. W. (2015). Are all biases missing data problems? *Current epidemiology reports*, 2(3):162–171.

Hughes, R. A., Davies, N. M., Davey Smith, G., and Tilling, K. (2019). Selection Bias When Estimating Average Treatment Effects Using One-sample Instrumental Variable Analysis. *Epidemiology*, 30(3):350–357.

Ihara, S. (1993). *Information Theory for Continuous Systems.* WORLD SCIENTIFIC.

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.

Infante-Rivard, C. and Cusson, A. (2018). Reflection on modern methods: selection bias—a review of recent developments. *International Journal of Epidemiology*, 47(5):1714–1722.

Joe, H. (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405):157–164.

Keele, L. J., Small, D. S., Hsu, J. Y., and Fogarty, C. B. (2019). Patterns of effects and sensitivity analysis for differences-in-differences.

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173.

Khazanchi, R., Evans, C. T., and Marcelin, J. R. (2020). Racism, Not Race, Drives Inequity Across the COVID-19 Continuum. *JAMA Network Open*, 3(9):e2019933–e2019933.

Kinney, J. B. and Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359.

Knox, D., Lowe, W., and Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637.

Kojadinovic, I. (2005). On the use of mutual information in data analysis : an overview.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions: Models and Applications*. Woley.

Krippendorff, K. (2009). Information of interactions in complex systems. *International Journal of General Systems*, 38(6):669–680.

Kumor, D., Cinelli, C., and Bareinboim, E. (2020). Efficient identification in linear structural causal models with auxiliary cutsets. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5501–5510. PMLR.

Laarne, P., Zaidan, M. A., and Nieminen, T. (2021). ennemi: Non-linear correlation detection with mutual information. *SoftwareX*, 14:100686.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620.

Larzelere, R. E., Kuhn, B. R., and Johnson, B. (2004). The intervention selection bias: An underrecognized confound in intervention research. *Psychological Bulletin*, 130(2):289–303.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., and Cole, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, 28(4):553–561.

Lett, E., Asabor, E., Beltrán, S., Cannon, A. M., and Arah, O. A. (2022). Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research. *The Annals of Family Medicine*, 20(2):157–163.

Linfoot, E. (1957). An informational measure of correlation. *Information and Control*, 1(1):85–89.

Lipsitch, M., Tchetgen, E. T., and Cohen, T. (2010). Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology*, 21(3):383–388.

Liu, R. A., Wei, Y., Qiu, X., Kosheleva, A., and Schwartz, J. D. (2022). Short term exposure to air pollution and mortality in the us: a double negative control analysis. *Environmental Health*, 21.

Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology*, 15(1).

Lu, H., Cole, S. R., Howe, C. J., and Westreich, D. (2022). Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology*, 33(5):699–706.

Lu, S. (2011). Measuring dependence via mutual information. *Master's thesis, Queen's University, Canada.*

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press.

Manjunath, B. G. and Wilhelm, S. (2021). Moments calculation for the doubly truncated multivariate normal density. *Journal of Behavioral Data Science*, 1(1):17–33.

Manski, C. F. and Pepper, J. V. (2018). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *The Review of Economics and Statistics*, 100(2):232–244.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7512–7523. PMLR.

Mathur, M. B. (2022). The M-Value: A Simple Sensitivity Analysis for Bias Due to Missing Data in Treatment Effect Estimates. *American Journal of Epidemiology*. kwac207.

Matthay, E. C. and Glymour, M. M. (2020). A graphical catalog of threats to validity. *Epidemiology*, 31(3):376–384.

McGill, W. (1954). Multivariate information transmission. *Psychometrika*, (19):97–116.

McMahan, I. (2017). Hacking the free throw: the science behind the most practiced shot in sports. https://www.theguardian.com/sport/2017/nov/22/free-throws-foul-shots-science-of-sports. [Online; posted 22-November-2017].

Meyer, P. E. (2014). *infotheo: Information-Theoretic Measures*. R package version 1.2.0.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Miao, W., Shi, X., and Tchetgen, E. T. (2020). A confounding bridge approach for double negative control inference on causal effects.

Michaud, I. (2018). *rmi: Mutual Information Estimators*. R package version 0.1.1.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

Mohan, K. and Pearl, J. (2014). Graphical models for recovering probabilistic and causal queries from missing data. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037.

Montgomery, J. M., Nyhan, B., and Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775.

Murphy, R. and Rohde, A. (2018). Rational bias in inflation expectations. *Eastern Economic Journal*, 44:153–171.

Necker, L. A. (1832). LXI. Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid.

Nguyen, T. Q., Dafoe, A., and Ogburn, E. L. (2019). The magnitude and direction of collider bias for binary variables. *Epidemiologic Methods*, 8(1):20170013.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan and Kaufmann, San Mateo.

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*, UAI'01, page 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). *Causality.* Cambridge University Press, Cambridge.

Pearl, J. (2014). The deductive approach to causal inference. *Journal of Causal Inference*, 2(2):115–129.

Pearl, J. (2015a). Conditioning on post-treatment variables. *Journal of Causal Inference*, 3(1):131–137.

Pearl, J. (2015b). Generalizing experimental findings. *Journal of Causal Inference*, 3(2):259–266.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, page 247–254.

Pearl, J. and Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579 – 595.

Pearl, J. and Bareinboim, E. (2019). Note on "generalizability of study results". *Epidemiology*, 30(2):186–188.

Rambachan, A. and Roth, J. (2021). An honest approach to parallel trends.

Richardson, T. and Robins, J. (2013a). Single world intervention graphs: a primer. *Working Paper, University of Washington, Seattle.*

Richardson, T. and Robins, J. (2013b). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Working Paper, Center for Statistics and the Social Sciences, University of Washington, Seattle*, (128).

Richiardi, L., Bellocco, R., and Zugna, D. (2013). Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*, 42(5):1511–1519.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Robinson, W. R. and Bailey, Z. D. (2019). Invited Commentary: What Social Epidemiology Brings to the Table—Reconciling Social Epidemiology and Causal Inference. *American Journal of Epidemiology*, 189(3):171–174.

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

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3):279–292.

Ryan, A. M., Kontopantelis, E., Linden, A., and James F Burgess, J. (2019). Now trending: Coping with non-parallel trends in difference-in-differences analysis. *Statistical Methods in Medical Research*, 28(12):3697–3711. PMID: 30474484.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10:441–451.

Saadati, M. and Tian, J. (2019). Adjustment criteria for recovering causal effects from missing data.

Schneider, E. B. (2020). Collider bias in economic history research. *Explorations in Economic History*, 78:101356.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston.

Shahar, D. J. and Shahar, E. (2017). A theorem at the core of colliding bias. *The International Journal of Biostatistics*, 13(1):20160055.

Sheehan, N., Didelez, V., Burton, P. R., and Tobin, M. D. (2008). Mendelian randomisation and causal inference in observational epidemiology. *PLoS medicine*, 5(8):e177.

Shevlyakov, G. and Vasilevskiy, N. (2017). A modification of linfoot's informational correlation coefficient. *Austrian Journal of Statistics*, 46(3-4):99–105.

Shi, X., Miao, W., and Tchetgen, E. (2020). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*, 7:190–202.

Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, page 352–359, Arlington, Virginia, USA. AUAI Press.

Shpitser, I., VanderWeele, T., and Robins, J. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*, Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010, pages 527–536. AUAI Press.

Sjölander, A. (2023). Selection bias with outcome-dependent sampling. *Epidemiology*, 34(2):186–191.

Smith, L. (2020). Selection mechanisms and their consequences: Understanding and addressing selection bias. *Current Epidemiology Reports*, 7:179–189.

Smith, L. H. and VanderWeele, T. J. (2019). Bounding bias due to selection. *Epidemiology*, 30(4):509–516.

Smith, R. (2015). A mutual information approach to calculating nonlinearity. *Stat*, 4(1):291–303.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen, E. J. T. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statist. Sci.*, 31(3):348–361.

Speed, T. (2011). A correlation for the 21st century. *Science*, 334(6062):1502–1503.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search.* MIT Press, Cambridge, MA.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465 – 472.

Stensrud, M. J., Hernán, M. A., Tchetgen Tchetgen, E. J., Robins, J. M., Didelez, V., and Young, J. G. (2021). A generalized theory of separable effects in competing event settings. *Lifetime Data Analysis*, 27(4):588–631.

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., and Hernán, M. A. (2022). Separable effects for causal inference in the presence of competing events. *Journal of the American Statistical Association*, 117(537):175–183.

Swanson, S. A. (2019). A Practical Guide to Selection Bias in Instrumental Variable Analyses. *Epidemiology*, 30(3):345–349.

Swanson, S. A. and Hernán, M. A. (2013). Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, 24(3):370–374.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., and Richardson, T. S. (2018). Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947. PMID: 31537952.

Swanson, S. A., Robins, J. M., Miller, M., and Hernán, M. A. (2015). Selecting on Treatment: A Pervasive Form of Bias in Instrumental Variable Analyses. *American Journal of Epidemiology*, 181(3):191–197.

Taddeo, M. M., Amorim, L. D., and Aquino, R. (2022). Causal measures using generalized difference-in-difference approach with nonlinear models. *Statistics and Its Interface*, 15(4):399 – 413.

Taleb, N. N. (2019). Fooled by correlation: Common misinterpretations in social "science".

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An introduction to proximal causal learning.

Tchetgen, E. T., Park, C., and Richardson, D. (2023a). Single proxy control.

Tchetgen, E. T., Park, C., and Richardson, D. (2023b). Universal difference-in-differences for causal inference in epidemiology.

Thompson, C. A. and Arah, O. A. (2014). Selection bias modeling using observed data augmented with imputed record-level probabilities. *Annals of Epidemiology*, 24(10):747–753.

Tripepi, G., Jager, K. J., Dekker, F. W., and Zoccali, C. (2010). Selection bias and information bias in clinical research. *Nephron Clinical Practice*, 115:94–99.

Van Der Zander, B. and Liśkiewicz, M. (2016). On searching for generalized instrumental variables. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1214–1222, Cadiz, Spain. PMLR.

Van Der Zander, B., Textor, J., and Liskiewicz, M. (2015). Efficiently finding conditional instruments for causal inference. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3243–3249. AAAI Press.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26.

VanderWeele, T. J. (2011). Controlled direct and mediated effects: definition, identification and bounds. *Scandinavian journal of statistics, theory and applications*, 38(3):551–563.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468.

Westreich, D. (2012). Berkson's bias, selection bias, and missing data. *Epidemiology*, 23(1):159–164.

Westreich, D., Edwards, J. K., Cole, S. R., Platt, R. W., Mumford, S. L., and Schisterman, E. F. (2015). Imputation approaches for potential outcomes in causal inference. *International Journal of Epidemiology*, 44(5):1731–1737.

Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R., and Stuart, E. A. (2018). Target Validity and the Hierarchy of Study Designs. *American Journal of Epidemiology*, 188(2):438–443.

Ye, T., Chen, S., and Zhang, B. (2022). The role of placebo samples in observational studies.

Ye, T., Keele, L., Hasegawa, R., and Small, D. S. (2020). A negative correlation strategy for bracketing in difference-in-differences.

Zhang, M. and Ding, P. (2022). Interpretable sensitivity analysis for the baron-kenny approach to mediation with unmeasured confounding.

Zhou, H. and Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4):493–504.