TECHNICAL REPORT
R-508
July 2021

UNIVERSITY OF CALIFORNIA
Los Angeles

Transparent and Robust Causal Inferences in the Social and Health Sciences

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

by

Carlos Leonardo Kulnig Cinelli

2021

ABSTRACT OF THE DISSERTATION

Transparent and Robust Causal Inferences
in the Social and Health Sciences

by

Carlos Leonardo Kulnig Cinelli
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2021
Professor Judea Pearl, Co-chair
Professor Chad J Hazlett, Co-chair

The past few decades have witnessed rapid and unprecedented theoretical progress on the science of causal inference. Most of this progress, however, relies on strong, *exact* assumptions, such as the absence of unobserved confounders, or the absence of certain direct effects. Unfortunately, more often than not these assumptions are not only untestable, but also very hard to defend in practice. This dissertation develops new theory, methods, and software for drawing causal inferences under more realistic settings. These tools allow applied scientists to both examine the sensitivity of their causal inferences to violations of their underlying assumptions, and also to draw robust (albeit also more modest) conclusions from settings in which traditional methods fail. Specifically, our contributions are the following: (i) novel powerful, yet simple, suite of sensitivity analysis tools for popular methods, such as confounding adjustment and instrumental variables, that can be immediately put to use to improve the robustness and transparency of current applied research; (ii) the first formal, systematic approach to sensitivity analysis for *arbitrary* linear structural causal models; and, (iii) novel (partial) identification results that marry two apparently disparate areas of causal inference research—the generalization of causal effects and the identification of "causes of effects." These methods are illustrated with examples from the social and health sciences.

The dissertation of Carlos Leonardo Kulnig Cinelli is approved.

Mark Stephen Handcock

Onyebuchi Aniweta Arah

Edward E Leamer

Judea Pearl, Committee Co-chair

Chad J Hazlett, Committee Co-chair

University of California, Los Angeles

2021

Contents

List of Figures

List of Tables

Acknowledgments

First of all, I am extremely grateful to my advisors, Chad Hazlett and Judea Pearl. Thank you very much for all the guidance and support throughout the PhD. If it is already exciting to witness a scientific revolution in the front row, it is even more exciting to be on the stage actively contributing to it, however modestly. That is how I feel about my experience here at UCLA, and this has been a great intellectual journey. I also would like to thank Mark Handcock, Onyebuchi Arah and Edward Leamer. Thank you not only for serving on this dissertation committee, but especially for all that I have learned with you, either in class, or in our informal meetings and conversations. Chapters 2 and 3 of this dissertation are based on [35] and [36] with Chad Hazlett (advisor), whereas Chapters 4 and 5 are based on [37] and [40] with Judea Pearl (advisor). Chapter 4 also has contributions from Bryant Chen, Daniel Kumor and Elias Bareinboim, who have helped with many discussions, text revisions and the extensive computational experiments using Gröbner basis. This research was partly supported by the UCLA Dissertation Year Fellowship (DYF), IBM Research, National Science Foundation [#IIS1704932], Office of Naval Research [#N00014-17-S-12091 and #N00014-21-1-2351] and Toyota Research Institute of North America [#PO-000897].

Curriculum Vitae

2010 – 2012	M.S. in Economics, University of Brasília (UnB).
2012 – 2016	Economist, Central Bank of Brazil.
2016 –	Ph.D. Student in Statistics, University of California, Los Angeles (UCLA).

Publications

[39] Carlos Cinelli and Judea Pearl. On the utility of causal diagrams in modeling attrition: a practical example. *Epidemiology*, 29(6):e50–e51, 2018.

[37] Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, and Elias Bareinboim. Sensitivity analysis of linear structural causal models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1252–1261. PMLR, 2019.

[11] Ioana Baldini, Clark W. Barrett, Antonio Chella, Carlos Cinelli, David Gamez, Leilani H. Gilpin, Knut Hinkelmann, Dylan Holmes, Takashi Kido, Murat Kocaoglu, William F. Lawless, Alessio Lomuscio, Jamie C. Macbeth, Andreas Martin, Ranjeev Mittu, Evan Patterson, Donald Sofge, Prasad Tadepalli, Keiki Takadama, and Shomir Wilson. Reports of the AAAI 2019 spring symposium series. *AI Mag.*, 40(3):59–66, 2019.

[35] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67,

2020.

[36] Carlos Cinelli and Chad Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Working Paper*, 2020.

[89] Daniel Kumor, Carlos Cinelli, and Elias Bareinboim. Efficient identification in linear structural causal models with auxiliary cutsets. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5501–5510. PMLR, 2020.

[38] Carlos Cinelli, Nathan LaPierre, Brian Hill, Sriram Sankararaman, and Eleazar Eskin. Robust mendelian randomization in the presence of residual population stratification, batch effects and horizontal pleiotropy. *bioRxiv*, 2020.

[31] Carlos Cinelli, Jeremy Ferwerda, and Chad Hazlett. sensemakr: Sensitivity analysis tools for OLS in R and Stata. *SSRN Electronic Journal*, Abstract ID 3588978, 2020.

[33] Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *SSRN Electronic Journal*, Abstract ID 3689437, 2020.

[147] Chi Zhang, Carlos Cinelli, Bryant Chen, and Judea Pearl. Exploiting equality constraints in causal inference. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1630–1638. PMLR, 2021.

[40] Carlos Cinelli and Judea Pearl. Generalizing experimental results by leveraging knowledge of mechanisms. *European Journal of Epidemiology*, 36:149 – 164, 2021.

CHAPTER 1

Introduction

Causal inference plays a vital role in the sciences, as it lies at the core of the most pressing issues facing society. How much racial bias is there in policing? Was this specific climate event due to global warming? Or, how many lives would have been saved from COVID-19, had mandatory mask use been implemented? These types of questions are impossible to answer from passive observations alone—no matter how big the data, or how sophisticated the machine learning system. To answer such questions, scientists need to rely on causal models. Toward this end, the past few decades have witnessed rapid and unprecedented theoretical progress on the science of causal inference, ranging from the popularization of certain "identification strategies" in the social and health sciences [6, 51], to the development of graphical causal models along with a complete solution to several non-parametric identification problems [109, 16]. These results have presented applied scientists with several conditions under which causal questions can be answered.

Most of this theoretical progress, however, relies on strong, *exact* assumptions about the data generating process, such as the absence of unobserved confounders (ignorability conditions), or the absence of certain direct effects (exclusion restrictions). Unfortunately, more often than not these assumptions are hard to defend in practice. This leads to two undesirable consequences for applied quantitative work: (i) important research questions may be neglected, simply because they do not exactly match the requirements of current methods; or, (ii) researchers may succumb to making the required "identification assumptions" (e.g, assuming ignorability of the treatment assignment or of the instrument) simply to justify the use of available methods, but not because these assumptions are truly believed (or understood). How much can we trust results based on doubtful—and often untestable—assumptions?

This dissertation develops new theory, methods, and software for drawing causal inferences under more realistic settings. These tools allow scientists, and policy makers to both examine the sensitivity of causal inferences to violations of its underlying assumptions, and also to draw robust (albeit also more modest) conclusions from settings in which traditional methods fail.[1] Specifically, our contributions are as follows.

Suite of sensitivity analysis tools for widely used methods (Chapters 2 and 3). The bulk of current applied work in causal inference with observational data still relies (perhaps unfortunately) on only a handful of identification results, namely: (i) adjusting for observed confounders (also known as "selection on observables"); (ii) instrumental variable methods; (iii) panel data methods (e.g, differences in differences, synthetic control); and, (iv) regression discontinuity designs. However, these methods require strong assumptions that most often are violated in practice. On the other hand, these assumptions need not hold precisely for an observational study to still be informative about the causal effect under investigation. In such cases, sensitivity analyses play an essential role, by allowing researchers to quantify how strong the violations of assumptions need to be in order to substantially change a research conclusion, and by aiding in determining whether such strong violations are plausible.

In Chapters 2 and 3, we develop flexible suite of sensitivity analysis tools fine-tuned for two of these widely used methods—confounding adjustment via regression models [35] and instrumental variable regression [36, 38]—along with accompanying software [31]. The benefits of these new methods are several. Not only are they conceptually easy to understand, but they also: (i) do not require any extra modeling assumptions; (ii) flexibly handle multiple or non-linear violations; (iii) exploit expert knowledge to bound the worst bias due to violations; and, (iv) are easy to compute using standard software. In particular, we introduce novel sensitivity statistics suited for *routine reporting*—such as the *robustness value*—describing

[1] For instance, it can be the case that we conclude the study is too fragile to plausible violations of its assumptions, and thus not sufficiently informative to answer the question it was supposed to address. An example is the study of Card [23] that uses "proximity to college" as an instrumental variable to investigate the causal effect of education on earnings. This is discussed on Section 3.5.

the minimum strength of violations needed to overturn the conclusions of a study. We also introduce the idea of *adjusted critical values*, in which researchers can easily perform inferences that are robust to systematic biases of any postulated strength, by simply replacing the usual critical threshold for a test statistic, (such as 1.96, for a 5% significance t-test) with a new, easy to compute adjusted threshold. Although recent, the results of these chapters have already been applied in several empirical studies across different fields, ranging from political science, economics, epidemiology and genetics.

An algorithmic approach to sensitivity analysis (Chapter 4). As discussed, currently there is a mismatch between the types of assumptions traditional causal inference is able to handle, and the types of assumptions scientists are willing to realistically defend. As current practices produce a steady stream of published results, it is important to handcraft the tools needed to bridge this gap for widely used models, as we do in Chapters 2 and 3. But going forward, we need to address the essence of this mismatch in a general way. This calls for a flexible, systematic approach to causal inference, that allows researchers to easily incorporate credible and realistic constraints in their models. For example, traditional identification results rely on exact assumptions about the absence of certain causal relationships. Our task is thus to *systematically* relax some of these assumptions, by allowing the possible presence of such relationships, albeit with limited strength.

In Chapter 4 we do this for the class of linear structural causal models. We develop an efficient, graph-based identification algorithm that *leverages non-zero constraints* on error covariances or path coefficient in arbitrary linear systems [37]. This technical result has several uses in itself (such as combining experimental data with observational data) but in particular it also allows the *algorithmic derivation of sensitivity curves*. Our results not only subsume several previous sensitivity analysis for canonical models (e.g, we can automatically derive sensitivity formulas for back-door adjustment, instrumental variable regression, front-door adjustment, mediation analysis—provided the relationships among variables are assumed to be linear), but it greatly expands the applicability of such type of analyses to many cases.

Generalizing experimental results, and causes of effects (Chapter 5). Two important areas of causal inference research are the generalization of causal effects across populations [113, 16, 78] and the identification of "causes of effects" [108, 137, 111, 112]. In Chapter 5 we show how these two apparently disparate areas of research can be merged for mutual benefit, unveiling important results in both areas [40]. We demonstrate how certain functional constraints may entail the invariance of *probabilities of causation* [108, 137] across domains, thus allowing the transport of causal effects (sometimes in the form of bounds) in settings where non-parametric generalization is otherwise impossible. The results of this chapter can thus be used both to transport (or bound) experimental findings from one population to another, as well as to quantify the percentage of individuals that are harmed by (or benefit from) a treatment, even when the average effect of this treatment is positive (or negative) in the population. These counterfactual probabilities may be important on their own right, and play a role in many applications of the social and health sciences, legal settings, and the production of explanations.

Chapter 6 concludes with some final remarks and directions for future work.

CHAPTER 2

Making Sense of Sensitivity: Extending Omitted Variable Bias

2.1 Introduction

Observational research often seeks to estimate causal effects under a “no unobserved confounding” or “ignorability” (conditional on observables) assumption (see e.g. 125, 109, 83). When making causal claims from observational data, investigators marshal what evidence they can to argue that their result is not due to confounding. In “natural” and “quasi”-experiments, this often includes a qualitative account for why the treatment assignment is “as-if” random conditional on a set of key characteristics (see e.g. 6, 51). Investigators seeking to make causal claims from observational data are also instructed to show “balance tests” and “placebo tests.” While, in some cases, null findings on these tests may be consistent with the claim of no unobserved confounders, they are certainly not dispositive: it is *unobserved* variables that we worry may be both “imbalanced” and related to the outcome in problematic ways. Fundamentally, causal inference always require assumptions that are unverifiable from the data [109].

Thus, in addition to balance and placebo tests, investigators are advised to conduct “sensitivity analyses” examining how fragile a result is against the possibility of unobserved confounding.[1] In general, such analyses entail two components: (1) describing the type of unobserved confounders—parameterized by their relation to the treatment assignment, the outcome, or both—that would substantively change our conclusions about the estimated causal

[1] Researchers may also wish to examine sensitivity to the choice of observed covariates, see [90, 91, 92].

effect, and (2) assisting the investigator in assessing the plausibility that such problematic confounding might exist, which necessarily depends upon the research design and expert knowledge regarding the data generating process. A variety of sensitivity analyses have been proposed, dating back to [42], with more recent contributions including [124, 118, 58, 120, 81, 21, 60, 76, 80, 141, 17, 61, 26, 50, 100, 104], and [62].

Yet, such sensitivity analyses remain underutilized.[2] We argue that a number of factors contribute to this reluctant uptake. One is the complicated nature and strong assumptions many of these methods impose, sometimes involving restrictions on or even a complete description of the nature of the confounder. A second reason is that, while training, convention and convenience dictate that users routinely report "regression tables" (or perhaps coefficient plots) to convey the results of a regression, we lack readily available quantities that aid in understanding and communicating how sensitive our results are to potential unobserved confounding. Third, and most fundamentally, connecting the results of a formal sensitivity analysis to a cogent argument about what types of confounders may exist in one's research project is often difficult, particularly with research designs that do not hinge on a credible argument regarding the (conditionally) "ignorable", "exogeneous", or "as-if random" nature of the treatment assignment. To complicate things, some of the solutions offered by the literature can lead users to erroneous conclusions (see Section 2.6 for discussion).

In this chapter we show how the familiar "omitted variable bias" (OVB) framework can be extended to address these challenges. We develop a suite of sensitivity analysis tools that do not require assumptions on the functional form of the treatment assignment mechanism nor on the distribution of the unobserved confounder, and can be used to assess the sensitivity to multiple confounders, whether they influence the treatment and outcome linearly or not.

We first introduce two novel measures of the sensitivity of linear regression coefficients: (i)

[2] In political science, out of 164 quantitative papers in the top three general interest publications (American Political Science Review, American Journal of Political Science, and Journal of Politics) for 2017, 64 papers clearly described a causal identification strategy other than a randomized experiment. Of these only 4 (6.25%) employed a formal sensitivity analyses beyond trying various specifications. In economics, [103] reports that most of non-experimental empirical papers utilized only informal robustness tests based on coefficient stability in the face of adding or dropping covariates. See also [29].

the "robustness value" (RV), which provides a convenient reference point to assess the overall robustness of a coefficient to unobserved confounding. If the confounders' association to the treatment and to the outcome (measured in terms of partial R^2) are *both* assumed to be less than the robustness value, then such confounders cannot "explain away" the observed effect. And, (ii) the proportion of variation in the outcome explained uniquely by the treatment, $R^2_{Y\sim D|\boldsymbol{X}}$, which reveals how strongly counfounders that explain 100% of the residual variance of the outcome would have to be associated with the treatment in order to eliminate the effect. Both measures can be easily computed from standard regression output: one needs only the estimate's t-value and the degrees of freedom. To advance standard practice across a variety of disciplines, we propose routinely reporting the RV and $R^2_{Y\sim D|\boldsymbol{X}}$ in regression tables.

Next, we offer graphical tools that investigators can use to refine their sensitivity analyses. The first is close in spirit to the proposal of [81]—a bivariate sensitivity contour plot, parameterizing the confounder in terms of partial R^2 values. However, contrary to Imbens' maximum likelihood approach, the OVB-based approach makes the underlying analysis simpler to understand, easier to compute, and more general. It side-steps assumptions on the functional form of the treatment assignment and on the distribution of the (possibly multiple, non-linear) confounders, and it easily extends contour plots to assess the sensitivity of t-values, p-values, or confidence intervals. This enables users to examine the types of confounders that would alter their inferential conclusions, not just point estimates. The second is an "extreme-scenario" sensitivity plot, in which investigators make conservative assumptions about the portion of otherwise unexplainend variance in the outcome that is due to confounders. One can then see how strongly such confounders would need to be associated with the treatment to be problematic. In the "worst-case" of these scenarios, the investigator assumes *all* unexplained variation in the outcome may be due to a confounder.

Finally, we introduce a novel bounding procedure that aids researchers in judging which confounders are plausible or could be ruled out, using the observed data in combination with expert knowledge. While prior work (58, 81, 76, 17, 50, 26, 100, 75) has suggested an informal practice of benchmarking the unobserved confounding by comparison to unadjusted statistics of observables, we show that this practice can generate misleading conclusions due to the effects of confounding itself, even if the confounder is assumed to be independent of the covariate(s) used for benchmarking. Instead, our approach formally bounds the strength of unobserved confounding with the same strength (or a multiple thereof) as a chosen observable or group of observables. These bounds are tight and may be especially useful when investigators can credibly argue to have measured the most important determinants (in terms of variance explained) of the treatment assignment or of the outcome.

In what follows, Section 2.2 describes the running example that will be used to illustrate the tools throughout the chapter—a study of the effect of violence on attitudes toward peace in Darfur, Sudan. Section 2.3 introduces the traditional OVB framework, how it can be used for a first approach to sensitivity analysis, and some of its shortcomings. Next, Section 2.4 shows how to extend the traditional OVB with the partial R^2 parameterization and Section 2.5 demonstrates how these results lead to a rich set of tools for sensitivity analysis. We conclude by discussing how our proposal seeks to increase the use of sensitivity analyses in practice, how it compares to existing procedures, and highlighting important caveats when interpreting sensitivity results. Open-source software for `R` and `Stata` implements the methods presented here.[3]

2.2 Running example

In this section we briefly introduce the applied example used throughout the chapter.[4] This serves as a background to illustrate how the tools developed here can be applied to address problems that commonly arise in observational research. We emphasize that the information produced by a sensitivity analysis is useful to the extent that researchers can wield domain knowledge about the data generating process to rule out the types of confounders shown to

[3] `R` package `sensemakr` [34] available on CRAN: `https://cran.r-project.org/package=sensemakr`. `Stata` module [32] available on SSC: `https://econpapers.repec.org/software/bocbocode/s458773.htm`. Web application available on: `https://carloscinelli.shinyapps.io/robustness_value/`. For details on how to use the software, we refer readers to [31].

[4] We only describe the most relevant details, further information is available in [72].

be problematic. Thus, a real world example helps to illustrate how such knowledge could be employed.

2.2.1 Exposure to violence in Darfur

In Sudan's western region of Darfur, a horrific campaign of violence against civilians began in 2003, sustaining high levels of violence through 2004, and killing an estimated 200,000 [55]. It was deemed genocide by then Secretary of State Colin Powell, and has resulted in indictments of alleged genocide, war crimes, and crimes against humanity in the International Criminal Court.

In the current case, we are interested in learning how being physically harmed during attacks on one's village changed individual attitudes towards peace. Clearly, we cannot randomize who is exposed to such violence. However, the means by which violence was distributed provide a tragic natural experiment. Violence against civilians during this time included both aerial bombardments by government aircraft, and attacks by a pro-government militia called the *Janjaweed*. While some villages were singled out for more or less violence, within a given village violence was arguably indiscriminate. This argument is supported by reports such as

> The government came with Antonovs, and targeted everything that moved. They made no distinction between the civilians and rebel groups. If it moved, it was bombed. It is the same thing, whether there are rebel groups (present) or not...The government bombs from the sky and the *Janjaweed* sweeps through and burns everything and loots the animals and spoils everything that they cannot take[5]

One can further argue that attacks were indiscriminate within village on the basis that the violence promoted by the government was mainly used to drive people out rather than target individuals. Within village, the bombing was crude and the attackers had almost no information about who they would target, with one major exception: while both men and

[5] Transcript from interview taken by Darfurian Voices team. Interview code 03072009_118_cf2009008.

women were often injured or killed, women were targeted for widespread sexual assault and rape by the *Janjaweed*.

With this in mind, an investigator might claim that village and gender are sufficient for control of confounding and estimate the linear model,

$$\text{PeaceIndex} = \hat{\tau}_{\text{res}}\text{DirectHarm} + \hat{\beta}_{f,\text{res}}\text{Female} + \text{Village}\hat{\boldsymbol{\beta}}_{v,\text{res}} + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{res}} + \hat{\varepsilon}_{\text{res}} \tag{2.1}$$

where *PeaceIndex* is an index measuring individual attitudes towards peace, *DirectHarm* a dummy variable indicating whether an individual was reportedly injured or maimed during such an attack, *Female* is a fixed effect for being female, and *Village* is a *matrix* of village fixed effects. Other pre-treatment covariates are included through the matrix $\boldsymbol{X}$, such as: age, whether they were a farmer, herder, merchant or trader, their household size and whether or not they voted in the past. The results of this regression show that, on average, exposure to violence (*DirectHarm*) is associated with more pro-peace attitudes on *PeaceIndex*.

Despite these arguments, not all investigators may agree with the assumption of no unobserved confounders. Consider, for example, a fellow researcher who argues that, although bombings were impossible to target finely, perhaps those in the center of the village were more often harmed than those on the periphery. And might not those nearer the center of each village also have different types of attitudes towards peace, on average? This suggests that the author ought to have instead run the model,

$$\text{PeaceIndex} = \hat{\tau}\text{DirectHarm} + \hat{\beta}_{f}\text{Female} + \text{Village}\hat{\boldsymbol{\beta}}_{v} + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\gamma}\text{Center} + \hat{\varepsilon}_{\text{full}} \tag{2.2}$$

That is, our earlier estimate $\hat{\tau}_{\text{res}}$ would differ from our target quantity $\hat{\tau}$. But how badly? How "strong" would a confounder like *Center* need to be to change our research conclusions? A simple violation of unconfoundedness such as this one can be handled in a relatively straightforward manner by the traditional OVB framework, as we will see in Section 2.3.

However, other skeptical researchers may question the claim that violence was conditionally indiscriminate with more elaborate stories, worrying that unobserved factors such as *Wealth* or *Political Attitudes* remain as confounders, perhaps even acting through non-linear functions such as an interaction of these two. Additionally, we may also have domain knowledge about the determinants of the outcome or the treatment assignment that could be used to limit arguments about potential confounding. For example, considering the nature of the attacks and the special role that gender played, one may argue that, within village, confounders are not likely to be as strongly associated with the treatment as the observed covariate *Female*.

How strong would these confounders need to be (acting as a group, possibly with non-linearities) to change our conclusions? And how could we *codify* and *leverage* our beliefs about the relative importance of *Female* to bound the plausible strength of unobserved confounders? In Sections 2.4 and 2.5, we show how extending the traditional OVB framework provides answers to such questions.

2.3 Sensitivity in an Omitted Variable Bias Framework

The "omitted variable bias" (OVB) formula is an important part of the mechanics of linear regression models and describes how the inclusion of an omitted covariate changes a coefficient estimate of interest. In this section, we review the traditional OVB approach, and illustrate its use as a simple tool for sensitivity analysis through bivariate contour plots showing how the effect estimate would vary depending upon hypothetical strengths of the confounder. This serves not only as an introduction to the method, but also to highlight limitations we will address in the following sections.

2.3.1 The traditional Omitted Variable Bias

Suppose an investigator wishes to run a linear regression model of an outcome Y on a treatment D, controlling for a set of covariates given by $\boldsymbol{X}$ and Z, as in

$$Y = \hat{\tau} D + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\gamma} Z + \hat{\varepsilon}_{\text{full}} \tag{2.3}$$

where Y is an $(n \times 1)$ vector containing the outcome of interest for each of the n observations and D is an $(n \times 1)$ treatment variable (which may be continuous or binary); $\boldsymbol{X}$ is an $(n \times p)$ matrix of *observed* (pre-treatment) covariates including the constant; and Z is a single $(n \times 1)$ *unobserved* covariate (we allow a multivariate version of Z in Section 2.4.5). However, since Z is unobserved, the investigator is forced instead to estimate a restricted model,

$$Y = \hat{\tau}_{\text{res}} D + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{res}} + \hat{\varepsilon}_{\text{res}} \tag{2.4}$$

where $\hat{\tau}_{\text{res}}$, $\hat{\boldsymbol{\beta}}_{\text{res}}$ are the coefficient estimates of the restricted OLS with only D and $\boldsymbol{X}$, *omitting* Z, and $\hat{\varepsilon}_{\text{res}}$ its corresponding residual.

How does the observed estimate $\hat{\tau}_{\text{res}}$ compare to the desired estimate, $\hat{\tau}$? Let us define as $\widehat{\text{bias}}$ the difference between these estimates, $\widehat{\text{bias}} := \hat{\tau}_{\text{res}} - \hat{\tau}$, where the hat, $\widehat{(\cdot)}$, clarifies that this quantity is a difference between sample estimates, not the difference between the expectation of a sample estimate and a population value. Using the Frisch-Waugh-Lovell (FWL) theorem (63, 93, 94) to "partial out" the observed covariates $\boldsymbol{X}$, the classic omitted variable bias solution is

$$\begin{aligned}
\hat{\tau}_{\text{res}} &= \frac{\text{cov}(D^{\perp \boldsymbol{X}},\ Y^{\perp \boldsymbol{X}})}{\text{var}(D^{\perp \boldsymbol{X}})} \\
&= \frac{\text{cov}(D^{\perp \boldsymbol{X}},\ \hat{\tau} D^{\perp \boldsymbol{X}} + \hat{\gamma} Z^{\perp \boldsymbol{X}})}{\text{var}(D^{\perp \boldsymbol{X}})} \\
&= \hat{\tau} + \hat{\gamma}\left(\frac{\text{cov}(D^{\perp \boldsymbol{X}},\ Z^{\perp \boldsymbol{X}})}{\text{var}(D^{\perp \boldsymbol{X}})}\right) \\
&= \hat{\tau} + \hat{\gamma}\hat{\delta}
\end{aligned} \tag{2.5}$$

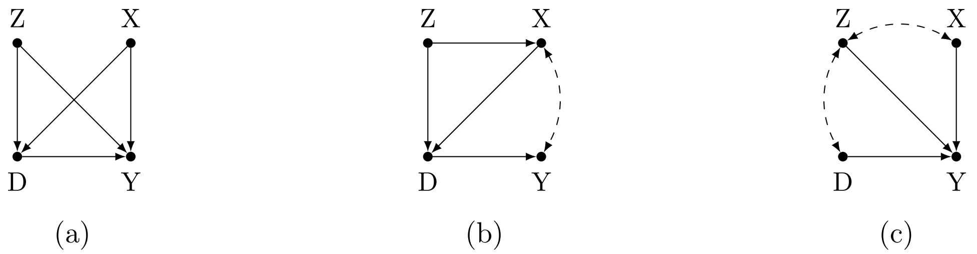
where $\text{cov}(\cdot)$ and $\text{var}(\cdot)$ denote the *sample* covariance and variance; $Y^{\perp \boldsymbol{X}}$, $D^{\perp \boldsymbol{X}}$ and $Z^{\perp \boldsymbol{X}}$ are the variables Y, D and Z after removing the components linearly explained by $\boldsymbol{X}$ and we define $\hat{\delta} := \frac{\text{cov}(D^{\perp \boldsymbol{X}}, Z^{\perp \boldsymbol{X}})}{\text{var}(D^{\perp \boldsymbol{X}})}$. We then have

$$\widehat{\text{bias}} = \hat{\gamma}\hat{\delta} \tag{2.6}$$

While elementary, the OVB formula in Equation 2.6 provides the key intuitions as well as a formulaic basis for a simple sensitivity analysis, letting us assess how the omission of covariates we wished to have controlled for could affect our inferences. Note that it holds *whether or not Equation 2.3 has a causal meaning.* In applied settings, however, one is typically interested in cases where the investigator has determined that the full regression, controlling for *both* $\boldsymbol{X}$ *and* the unobserved variable Z, would have identified the causal effect of D on Y; thus, hereafter we will treat Z as an unobserved "confounder" and continue the discussion as if the estimate $\hat{\tau}$, obtained with the inclusion of Z, is the desired target quantity.

A note on identification via covariate adjustment

Conditions that endow regression estimates with causal meaning are extensively discussed in the literature: identification assumptions can be articulated in graphical terms, such as postulating a structural causal model in which $\{\boldsymbol{X}, Z\}$ satisfy the backdoor or adjustment criterion for identifying the causal effect of D on Y [109, 127]; or, equivalently, in counterfactual notation, stating that the treatment assignment D is conditionally ignorable given $\{\boldsymbol{X}, Z\}$, that is $Y_d \perp\!\!\!\perp D | \{\boldsymbol{X}, Z\}$, where Y_d denotes the potential outcome of Y when D is *set* to d (see 109, 6, 127, 83). To illustrate, Figure 2.1 shows the causal diagrams of three distinct models in which the set $\{X, Z\}$ satisfy the backdoor criterion for identifying the causal effect of D on Y, and thus conditional ignorability, $Y_d \perp\!\!\!\perp D \mid \{X, Z\}$, holds. We further note the effect of D on Y may be non-linear, in which case a regression coefficient may be an incomplete summary of the causal effect of interest [6]. Finally, indiscriminate inclusion of covariates can induce or amplify bias (see [110, 47, 100, 134] for related discussions; see also [33] for a visual summary of simple graphical criteria to distinguish "good" from "bad" controls, both for deciding which set of variables should be adjusted for to identify the causal effect of interest, as well as deciding which, among a set of valid adjustment sets, would yield more precise estimates). Here we assume the researcher is interested in the estimates one would obtain from running the regression in Equation 2.3, controlling for both observed variables $\boldsymbol{X}$ and an unobserved variable Z (generalization to a multivariate $\boldsymbol{Z}$ is given in Section 2.4.5).



(a) (b) (c)

Figure 2.1: Different causal diagrams in which the set $\{X, Z\}$ satisfy the backdoor criterion for identifying the causal effect of D on Y and thus $Y_d \perp\!\!\!\perp D \mid \{X, Z\}$ holds. Dashed bi-directed edges stand for latent confounders between variables.

2.3.2 Making sense of the traditional OVB

One virtue of the OVB formula is its interpretability. The quantity $\hat{\gamma}$ describes the difference in the linear expectation of the outcome, when comparing individuals that differ by one unit on the confounder, but have the same treatment assignment status as well as the same value for all remaining covariates. In broader terms, $\hat{\gamma}$ describes how looking at different subgroups of the unobserved confounder "impacts" our best linear prediction of the outcome.[6]

By analogy, it would be tempting to think of $\hat{\delta}$ as the estimated marginal "impact" of the confounder on the *treatment*. However, causal interpretation aside, this is incorrect because it refers instead to the coefficient of the reverse regression, $Z = \hat{\delta} D + \boldsymbol{X}\hat{\boldsymbol{\psi}} + \hat{\varepsilon}_Z$, and not the regression of the treatment D on Z, and $\boldsymbol{X}$. That is, $\hat{\delta}$ gives the difference in the linear expectation of the confounder, when comparing individuals with the same values for the covariates, but differing by one unit on the treatment. This quantity will be familiar to empirical researchers who have used quasi-experiments in which the treatment is believed to

[6] While a causal interpretation here is tempting, whether this difference in the distribution of the outcome within strata of the confounder can be attributed to a direct causal effect of the former on the latter depends on structural assumptions. Suppose, for example, the "true" outcome model is assumed to be a linear structural equation where strict exogeneity holds, i.e., $Y = \tau D + \boldsymbol{X}\boldsymbol{\beta} + \gamma Z + \varepsilon$ and $\mathbb{E}[\varepsilon | D, \boldsymbol{X}, Z] = 0$. Then, $\hat{\gamma}$ could be interpreted as an estimate of the direct causal impact of a unit change of the confounder on the expected value of the outcome Y, holding the other covariates fixed. In many scenarios, however, this is unrealistic—since the researcher's goal is to estimate the causal effect of D on Y, usually Z is required only to, along with $\boldsymbol{X}$, block the back-door paths from D to Y [109], or equivalently, make the treatment assignment conditionally ignorable. In this case, $\hat{\gamma}$ could reflect not only its causal effect on Y (if it has any) but also other spurious associations not eliminated by standard assumptions. One such example is provided by the causal diagram of Figure 2.1b. Heuristically, however, referring to $\hat{\gamma}$ as the marginal "impact" of the confounder on the outcome is useful, as long as the reader keeps in mind that it is an associational quantity with causal meaning only under certain circumstances.

be randomized only conditional on certain covariates $\boldsymbol{X}$. In that case we may then check for "balance" on other (pre-treatment) observables once conditioning is complete. Hence, we can think of $\hat{\delta}$ as the (conditional) imbalance of the confounder with respect to the treatment—or simply "imbalance".

Thus, a useful mnemonic is that the omitted variable bias can be summarized as the unobserved confounder's "impact times its imbalance". Note that the imbalance component is quite general: whatever the true functional form dictating $\mathbf{E}[Z|D, \boldsymbol{X}]$ (or the treatment assignment mechanism), the only way in which Z's relationship to D enters the bias is captured by its "linear imbalance", parameterized by $\hat{\delta}$. In other words, the linear regression of Z on D and $\boldsymbol{X}$ need not reflect the correct expected value of Z—rather it serves to capture the aspects of the relationship between Z and D that affects the bias.

2.3.3 Using the traditional OVB for sensitivity analysis

If we know the *signs* of the partial correlations between the confounder with the treatment and the outcome (the same as the signs of $\hat{\gamma}$ and $\hat{\delta}$) we can argue whether our estimate is likely to be underestimating or overestimating the quantity of interest. Arguments using correlational direction is common practice in econometrics work.[7] Often, though, discussing possible direction of the bias is not possible or not sufficient, and magnitude must be considered. How strong would the confounder(s) have to be to change the estimates in such a way to affect the main conclusions of a study?

Sensitivity contour plots

A first approach to investigate the sensitivity of our estimate can be summarized by a two-dimensional plot of bias contours parameterized by the two terms $\hat{\gamma}$ and $\hat{\delta}$. Each pair of hypothesized "impact" and "imbalance" parameters corresponds to a certain level of bias

[7] e.g. "Using a similar omitted-variables-type argument, we note that even if there are other confounders that we haven't controlled for, those that are positively correlated with private school attendance are likely to be positively correlated with earnings as well. Even if these variables remain omitted, their omission leads the estimates computed with the variables at hand to overestimate the private school premium." [8, p.8-9]

(their product), but given an initial treatment effect estimate $\hat{\tau}_{\text{res}}$, we can also relabel the bias levels in terms of the "adjusted" effect estimate, i.e $\hat{\tau} = \hat{\tau}_{\text{res}} - \hat{\gamma}\hat{\delta}$, the estimate from the OLS regression we wish we had run, if we had included a confounder with the hypothesized level of impact and imbalance.

In our running example, a specific confounder we wish we had controlled for is a binary indicator of whether the respondent lived in the center or in the periphery of the village. How strong would this specific confounder have to be in order for its inclusion to substantially affect our conclusions? Figure 2.2 shows the plot of adjusted estimates for several hypothetical values of impact and imbalance of the confounder *Center*.

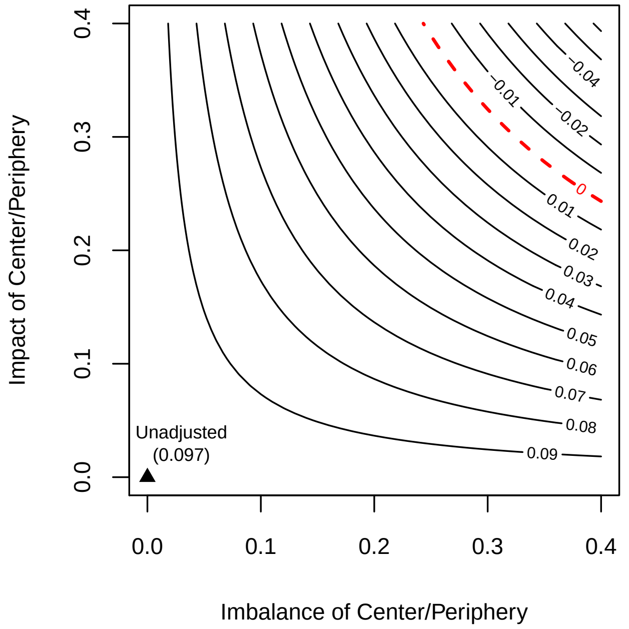


Figure 2.2: Sensitivity contours of the point estimate—traditional OVB.

Hypothetical values for the imbalance of the confounder lie on the horizontal axis. In this particular case, they indicate how those who were harmed are hypothesized to differ from those who were not harmed in terms of the proportion of people living in the center of the village. Values for the hypothetical impact of the confounder on the outcome lie on the vertical axis, representing how attitudes towards peace differ on average for people living in the center versus those in the periphery of the village, within strata of other covariates.

The contour lines of the plot give the adjusted treatment effect at hypothesized values of the impact and imbalance parameters. They show the exact estimate one would have obtained by running the full regression including a confounder with those hypothetical sensitivity parameters. No other information is required to know how such a confounder would influence the result. Notice that here, and throughout the chapter, we parameterize the bias in a way that it hurts our preferred hypothesis by reducing the absolute effect size.[8]

This plot explicitly reveals the type of prior knowledge one needs to have in order to be able to rule out problematic confounders. As an example, imagine the confounder *Center* has a conditional imbalance as high as 0.25—that is, having controlled for the observed covariates, those who were physically injured were also 25 percentage points more likely to live in the center of the village than those who were not. With such an imbalance, the plot reveals that the impact of living in the center on the outcome (Peace Index) would have to be over 0.40 in order to bring down the estimated effect of *DirectHarm* to zero.

Determining whether this is good or bad news remains difficult and requires contextual knowledge about the process that generated the data. For instance, one could argue that, given the relatively homogeneous nature of these small villages and that their centers are generally not markedly different in composition than the peripheries, it is hard to believe that being in the center was associated with a 0.40 higher expected score on Peace Index (which varies only from 0 to 1). Regardless of whether the investigator can make a clear argument that rules out such confounders, the virtue of sensitivity analysis is that it moves the conversation from one where the investigator seeks to defend "perfect identification" and the critic points out potential confounders, to one where details can be given and discussed about the degree of confounding that would be problematic.

[8] Investigators may also argue that accounting for omitted variable bias would increase the effect size, in the sense that the current estimate is conservative. Our tools apply to these cases as well, the arguments would just work in the opposite direction. For simplicity of exposition, in the chapter we focus on the case where accounting for omitted variable bias reduces the effect size.

Shortcomings of the traditional OVB

The traditional OVB has some benefits: as shown, with sound substantive knowledge about the problem, it is a straightforward exercise. But it also has shortcomings. In the previous example, *Center* was a convenient choice of confounder because it is a binary variable, and the units of measure attached to "impact" and "imbalance" are thus easy to understand as changes in proportions. This is not in general the case. Imagine contemplating confounders such as *Political Attitudes*: in what scale should we measure this? A doubling of that scale would halve the required "impact" and double the required "imbalance". A possible solution is standardizing the coefficients, but this does not help if the goal is to assess the sensitivity of the causal parameter in its original scale.

Furthermore, the traditional OVB, be it standardized or not, does not generalize easily to multiple confounders: how should we assess the effect of confounders *Political Attitudes* and *Wealth*, acting together, perhaps with complex non-linearities? Or, more generally, how should we consider all the other unnamed confounders acting together? Can we benchmark all these confounders against *Female*? Finally, how can we obtain the sensitivity of not only the point estimate, but also the standard errors, so that we could examine t-values, p-values or confidence intervals under hypothetical confounders?

2.4 OVB with the partial R^2 parameterization

We now consider a reparameterization of the OVB formula in terms of partial R^2 values. Our goal is to replace the sensitivity parameters $\hat{\gamma}$ and $\hat{\delta}$ with a pair of parameters that uses an R^2 measure to assess the strength of association between the confounder and the treatment and between the confounder and the outcome, both assuming the remaining covariates $\boldsymbol{X}$ have been accounted for. The partial R^2 parameterization is scale-free and it further enables us to construct a number of useful analyses, including: (i) assessing the sensitivity of an estimate to any number or even *all* confounders acting together, possibly non-linearly; (ii) using the same framework to assess the sensitivity of point estimates as well as t-values and confidence intervals; (iii) assessing the sensitivity to extreme-scenarios in which all or a big portion of the unexplained variance of the outcome is due to confounding; (iv) applying contextual information about the research design to bound the strength of the confounders; and (v) presenting these sensitivity results concisely for easy routine reporting, as well as providing visual tools for finer grained analysis.

2.4.1 Reparameterizing the bias in terms of partial R^2

Let $R^2_{Z\sim D}$ denote the (sample) R^2 of regressing Z on D. Recall that for OLS the following holds, $R^2_{Z\sim D} = \frac{\text{var}(\hat{Z})}{\text{var}(Z)} = 1 - \frac{\text{var}(Z^{\perp D})}{\text{var}(Z)} = \text{cor}(Z, \hat{Z})^2 = \text{cor}(Z, D)^2$, where $\hat{Z}$ are the fitted values given by regressing Z on D. Notice the R^2 is symmetric, that is, it is invariant to whether one uses the "forward" or the "reverse" regression since $R^2_{Z\sim D} = \text{cor}(Z, D)^2 = \text{cor}(D, Z)^2 = R^2_{D\sim Z}$. Extending this to the case with covariates $\boldsymbol{X}$, we denote the partial R^2 from regressing Z on D after controlling for $\boldsymbol{X}$ as $R^2_{Z\sim D|\boldsymbol{X}}$. This has the same useful symmetry, with $R^2_{Z\sim D|\boldsymbol{X}} = 1 - \frac{\text{var}(Z^{\perp \boldsymbol{X},D})}{\text{var}(Z^{\perp \boldsymbol{X}})} = \text{cor}(Z^{\perp \boldsymbol{X}}, D^{\perp \boldsymbol{X}})^2 = \text{cor}(D^{\perp \boldsymbol{X}}, Z^{\perp \boldsymbol{X}})^2 = R^2_{D\sim Z|\boldsymbol{X}}$.

We are now ready to express the bias in terms of partial R^2. First, by the FWL theorem,

$$\begin{aligned}
\widehat{\text{bias}} &= \hat{\delta}\hat{\gamma} \\
&= \left(\frac{\text{cov}(D^{\perp \boldsymbol{X}},\ Z^{\perp \boldsymbol{X}})}{\text{var}(D^{\perp \boldsymbol{X}})}\right)\left(\frac{\text{cov}(Y^{\perp \boldsymbol{X},D},\ Z^{\perp \boldsymbol{X},D})}{\text{var}(Z^{\perp \boldsymbol{X},D})}\right) \\
&= \left(\frac{\text{cor}(D^{\perp \boldsymbol{X}},\ Z^{\perp \boldsymbol{X}})\text{sd}(Z^{\perp \boldsymbol{X}})}{\text{sd}(D^{\perp \boldsymbol{X}})}\right)\left(\frac{\text{cor}(Y^{\perp \boldsymbol{X},D},\ Z^{\perp \boldsymbol{X},D})\text{sd}(Y^{\perp \boldsymbol{X},D})}{\text{sd}(Z^{\perp \boldsymbol{X},D})}\right) \\
&= \left(\frac{\text{cor}(Y^{\perp \boldsymbol{X},D},\ Z^{\perp \boldsymbol{X},D})\text{cor}(D^{\perp \boldsymbol{X}},\ Z^{\perp \boldsymbol{X}})}{\frac{\text{sd}(Z^{\perp \boldsymbol{X},D})}{\text{sd}(Z^{\perp \boldsymbol{X}})}}\right)\left(\frac{\text{sd}(Y^{\perp \boldsymbol{X},D})}{\text{sd}(D^{\perp \boldsymbol{X}})}\right)
\end{aligned} \tag{2.7}$$

Noting that $\text{cor}(Y^{\perp \boldsymbol{X},D}, Z^{\perp \boldsymbol{X},D})^2 = R^2_{Y\sim Z|\boldsymbol{X},D}$, that $\text{cor}(Z^{\perp \boldsymbol{X}},\ D^{\perp \boldsymbol{X}})^2 = R^2_{D\sim Z|\boldsymbol{X}}$, and that $\frac{\text{var}(Z^{\perp \boldsymbol{X},D})}{\text{var}(Z^{\perp \boldsymbol{X}})} = 1 - R^2_{Z\sim D|\boldsymbol{X}} = 1 - R^2_{D\sim Z|\boldsymbol{X}}$, we can write 2.7 as

$$|\widehat{\text{bias}}| = \sqrt{\frac{R^2_{Y\sim Z|D,\boldsymbol{X}}\ R^2_{D\sim Z|\boldsymbol{X}}}{1 - R^2_{D\sim Z|\boldsymbol{X}}}}\left(\frac{\text{sd}(Y^{\perp \boldsymbol{X},D})}{\text{sd}(D^{\perp \boldsymbol{X}})}\right). \tag{2.8}$$

Equation 2.8 rewrites the OVB formula in terms that more conveniently rely on partial R^2 measures of association rather than raw regression coefficients. Investigators may be interested in how confounders alter inference as well, so we also examine the standard error. Let df denote the regression's degrees of freedom (for the restricted regression actually run). Noting that

$$\widehat{\text{se}}(\hat{\tau}_{\text{res}}) = \frac{\text{sd}(Y^{\perp \boldsymbol{X},D})}{\text{sd}(D^{\perp \boldsymbol{X}})}\sqrt{\frac{1}{\text{df}}} \tag{2.9}$$

$$\widehat{\text{se}}(\hat{\tau}) = \frac{\text{sd}(Y^{\perp \boldsymbol{X},D,Z})}{\text{sd}(D^{\perp \boldsymbol{X},Z})}\sqrt{\frac{1}{\text{df}-1}}, \tag{2.10}$$

whose ratio is

$$\frac{\widehat{\text{se}}(\hat{\tau})}{\widehat{\text{se}}(\hat{\tau}_{\text{res}})} = \left(\frac{\text{sd}(Y^{\perp \boldsymbol{X},D,Z})}{\text{sd}(Y^{\perp \boldsymbol{X},D})}\right)\left(\frac{\text{sd}(D^{\perp \boldsymbol{X}})}{\text{sd}(D^{\perp \boldsymbol{X},Z})}\right)\sqrt{\frac{\text{df}}{\text{df}-1}}, \tag{2.11}$$

we obtain the expression for the estimated standard error of $\hat{\tau}$

$$\widehat{\text{se}}(\hat{\tau}) = \widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\frac{1-R^2_{Y\sim Z|D,\boldsymbol{X}}}{1-R^2_{D\sim Z|\boldsymbol{X}}}\left(\frac{\text{df}}{\text{df}-1}\right)}. \tag{2.12}$$

Moreover, with this we can further see the bias as

$$|\widehat{\text{bias}}| = \widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\frac{R^2_{Y\sim Z|D,\boldsymbol{X}}\; R^2_{D\sim Z|\boldsymbol{X}}}{1-R^2_{D\sim Z|\boldsymbol{X}}}(\text{df})}. \tag{2.13}$$

2.4.2 Making sense of the partial R^2 parameterization

Equations 2.12 and 2.13 form the basis of the sensitivity exercises regarding both the point estimate and the standard error, with sensitivity parameters in terms of $R^2_{Y\sim Z|D,\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$. These formulae are computationally convenient—the only data dependent parts are the standard error of $\hat{\tau}_{\text{res}}$ and the regression's degrees of freedom, which are already reported by most regression software. In this section, we provide remarks that help making sense of these results, revealing their simplicity in terms of regression anatomy. We also review some partial R^2 identities that may prove useful when reasoning about the sensitivity parameters.

Sensitivity of the point estimate

In the partial R^2 parameterization, the relative bias, $\left|\frac{\widehat{\text{bias}}}{\hat{\tau}_{\text{res}}}\right|$, has a simple form:[9]

$$\text{relative bias} = \frac{\overbrace{|R_{Y\sim Z|D,\boldsymbol{X}} \times f_{D\sim Z|\boldsymbol{X}}|}^{\text{bias factor}}}{\underbrace{|f_{Y\sim D|\boldsymbol{X}}|}_{\text{partial f of D with Y}}} = \frac{\text{BF}}{|f_{Y\sim D|\boldsymbol{X}}|}. \tag{2.14}$$

The numerator of the relative bias contains the partial Cohen's f of the confounder with the treatment, "amortized" by the partial correlation of that confounder with the outcome.[10] Collectively this numerator could be called the "bias factor" of the confounder, $\text{BF} = |R_{Y\sim Z|D,\boldsymbol{X}} \times f_{D\sim Z|\boldsymbol{X}}|$, which is determined entirely by the two sensitivity parameters $R^2_{Y\sim Z|D,\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$. To determine the size of the relative bias, this is compared to how much variation of the outcome is uniquely explained by the treatment assignment, in the form of the partial Cohen's f of the treatment with the outcome. Computationally, $f_{Y\sim D|\boldsymbol{X}}$ can be obtained by dividing the t-value of the treatment coefficient by the square-root of the regression's degrees of freedom—$f_{Y\sim D|\boldsymbol{X}} = t_{\hat{\tau}_{\text{res}}}/\sqrt{\text{df}}$. This allows one to easily assess sensitivity to any confounder with a given pair of partial R^2 values.

Equation 2.14 also reveals that, given a particular confounder (which will fix BF), the only property needed to determine the robustness of a regression estimate against that confounder is the partial R^2 of the treatment with the outcome (via $f_{Y\sim D|\boldsymbol{X}}$). This serves to reinforce the fact that robustness to confounding is an identification problem, impervious to sample size considerations. While t-values and p-values might be informative with respect to the statistical uncertainty (in a correctly specified model), robustness to misspecification is determined by the share of variation of the outcome the treatment uniquely explains.

[9] See appendix for details.

[10] Cohen's f^2 can be written as $f^2 = R^2/(1-R^2)$, so, for example, $f^2_{D\sim Z|\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}}/(1-R^2_{D\sim Z|\boldsymbol{X}})$.

A subtle but useful property of the partial R^2 parameterization is that it reveals an asymmetry in the role of the components of the bias factor. In the traditional OVB formulation, the bias is simply a product of two terms with the same importance. The new formulation breaks this symmetry: the effect of the partial R^2 of the confounder with the outcome on the bias factor is bounded at one. By contrast, the effect of the partial R^2 of the confounder with the treatment on the bias factor is unbounded (via $f_{D\sim Z|\boldsymbol{X}}$). This allows us to consider extreme scenarios, in which we suppose the confounder explains *all* of the left-out variation of the outcome, and see what happens as we vary the partial R^2 of the confounder with the treatment (Section 2.5.3).

Sensitivity of the variance

How the confounder affects the estimate of the variance has a straightforward interpretation as well. The relative change in the variance, $\frac{\widehat{\text{var}}(\hat{\tau})}{\widehat{\text{var}}(\hat{\tau}_{\text{res}})}$, can be decomposed into three components,

$$\begin{aligned}\text{relative change in variance} &= \overbrace{\left(1 - R^2_{Y\sim Z|D,\boldsymbol{X}}\right)}^{\text{VRF}} \underbrace{\left(\frac{1}{1 - R^2_{D\sim Z|\boldsymbol{X}}}\right)}_{\text{VIF}} \overbrace{\left(\frac{\text{df}}{\text{df} - 1}\right)}^{\text{change in df}} \\ &= \text{VRF} \times \text{VIF} \times \text{change in df.} \qquad (2.15)\end{aligned}$$

That is, including the confounder in the regression reduces the estimate of the variance of the coefficient of D by reducing the residual variance of Y (variance reduction factor—VRF). On the other hand, it raises the estimated variance of the coefficient via its partial correlation with the treatment (the traditional variance inflation factor—VIF). Finally, the degrees of freedom must be adjusted to formally recover the answer one would obtain from including the omitted variable. The overall relative change of the estimated variance is simply the product of these three components.

Reasoning about $R^2_{Y\sim Z|D,\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$

For simplicity of exposition, throughout the chapter we reason in terms of the sensitivity parameters $R^2_{Y\sim Z|D,\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$ directly. However, here we recall some identities of the partial R^2 scale that can aid interpretation depending upon what can best be reasoned about in a given applied setting.

First, as noted in Section 2.4.1, researchers accustomed to thinking about or evaluating the strength of (partial) correlations can simply square those values to reason with the corresponding partial R^2s. Next, in some circumstances, researchers might prefer to reason about the relationship of the unobserved confounder Z and the outcome Y *without conditioning on the treatment assignment* D.[11] This can be done by noting that, for a choice of $R_{Y\sim Z|\boldsymbol{X}}$ and $R_{D\sim Z|\boldsymbol{X}}$, we can reconstruct $R_{Y\sim Z|D,\boldsymbol{X}}$ using the recursive definition of partial correlations,

$$R_{Y\sim Z|D,\boldsymbol{X}} = \frac{R_{Y\sim Z|\boldsymbol{X}} - R_{Y\sim D|\boldsymbol{X}} R_{D\sim Z|\boldsymbol{X}}}{\sqrt{1-R^2_{Y\sim D|\boldsymbol{X}}}\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}}}}. \tag{2.16}$$

Therefore, if needed, one can reason directly about sensitivity parameters $R^2_{Y\sim Z|\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$.

Finally, it may be beneficial to reason in terms of how much explanatory power is added by including confounders. To this end, recall the partial R^2's are defined as,

$$R^2_{Y\sim Z|D,\boldsymbol{X}} = \frac{R^2_{Y\sim D+\boldsymbol{X}+Z} - R^2_{Y\sim D+\boldsymbol{X}}}{1-R^2_{Y\sim D+\boldsymbol{X}}}, \qquad R^2_{D\sim Z|\boldsymbol{X}} = \frac{R^2_{D\sim \boldsymbol{X}+Z} - R^2_{D\sim \boldsymbol{X}}}{1-R^2_{D\sim \boldsymbol{X}}}. \tag{2.17}$$

That is, plausibility judgments about the partial R^2 boil down to plausibility judgments about the *total (or added) explanatory power* that one would have obtained in the treatment and the outcome regressions, had the unobserved confounder Z been included. This may be particularly useful when contemplating multiple confounders acting in concert (as we will

[11] For instance, since D will usually be a *post-treatment* variable with respect to Z, this can make the association of Y and Z conditional on D harder to interpret, especially when one wants to attach a causal meaning to the parameter [119]. As argued in footnote 6, however, recall that a causal interpretation of the association of Z with Y requires more assumptions than the ones usually invoked for the identification of the causal effect of D on Y.

discuss in Section 2.4.5), in which case other parameterizations (such as simple correlations or regression coefficients) become unwieldy.

2.4.3 Sensitivity statistics for routine reporting

Detailed sensitivity analyses can be conducted using the previous results, as we will show in the next section. However, widespread adoption of sensitivity analyses would benefit from simple measures that quickly describe the overall sensitivity of an estimate to unobserved confounding. These measures serve two main purposes: (i) they can be routinely reported in standard regression tables, making the discussion of sensitivity to unobserved confounding more accessible and standardized; and, (ii) they can be easily computed from quantities found on a regression table, allowing readers and reviewers to initiate the discussion about unobserved confounders when reading papers that did not formally assess sensitivity.

The robustness value

The first quantity we propose is the *robustness value* (RV), which conveniently summarizes the types of confounders that would problematically change the research conclusions. Consider a confounder with equal association to the treatment and the outcome, i.e. $R^2_{Y\sim Z|\boldsymbol{X},D} = R^2_{D\sim Z|\boldsymbol{X}} = \text{RV}_{q^*}$. The RV_{q^*} describes how strong that association must be in order to reduce the estimated effect by $(100 \times q^*)\%$. By Equation 2.14 (see appendix 7.1.1),

$$\text{RV}_{q^*} = \frac{1}{2}\left(\sqrt{f^4_{q^*} + 4f^2_{q^*}} - f^2_{q^*}\right) \tag{2.18}$$

where $f_{q^*} := q^*|f_{Y\sim D|\boldsymbol{X}}|$ is the partial Cohen's f of the treatment with the outcome multiplied by the proportion of reduction q^* on the treatment coefficient which would be deemed problematic. Confounders that explain $(100 \times \text{RV}_{q^*})\%$ of the residual variance both of the treatment and the outcome are sufficiently strong to change the point estimate in problematic ways, while confounders with neither association greater than $(100 \times \text{RV}_{q^*})\%$ are not.

The RV thus offers an interpretable sensitivity measure that summarizes how robust the point estimate is to unobserved confounding. A robustness value close to one means the treatment effect can handle strong confounders explaining almost all residual variation of the treatment and the outcome. On the other hand, a robustness value close to zero means that even very weak confounders could eliminate the results. Note that the RV can be easily computed from any regression table, recalling that $f_{Y\sim D|\boldsymbol{X}}$ can be obtained by simply dividing the treatment coefficient t-value by $\sqrt{\text{df}}$.

With minor adjustment, robustness values can also be obtained for t-values, or lower and upper bounds of confidence intervals. Let $|t^*_{\alpha,\text{df}-1}|$ denote the t-value threshold for a t-test with significance level of α and $\text{df}-1$ degrees of freedom, and define $f^*_{\alpha,\text{df}-1} := |t^*_{\alpha,\text{df}-1}|/\sqrt{\text{df}-1}$. Now construct an adjusted $f_{q^*,\alpha}$, accounting for both the proportion of reduction q^* of the point estimate and the boundary below which statistical significance is lost at the level of α,

$$f_{q^*,\alpha} := f_{q^*} - f^*_{\alpha,\text{df}-1} \tag{2.19}$$

If $f_{q^*,\alpha} < 0$, then the robustness value is zero. If $f_{q^*,\alpha} > 0$, then a confounder with a partial R^2 of,

$$\text{RV}_{q^*,\alpha} = \frac{1}{2}\left(\sqrt{f^4_{q^*,\alpha} + 4f^2_{q^*,\alpha}} - f^2_{q^*,\alpha}\right), \tag{2.20}$$

both with the treatment and with the outcome is sufficiently strong to make the adjusted t-test not reject the null hypothesis $H_0 : \tau = (1-q^*)|\hat{\tau}_{\text{res}}|$ at the α level (or, equivalently, to make the adjusted $1-\alpha$ confidence interval include $(1-q^*)|\hat{\tau}_{\text{res}}|$). When $\text{RV}_{q^*,\alpha} > 1 - 1/f^2_{q^*}$ then, as with the RV_{q^*}, we can conclude that no confounder with both associations lower than $\text{RV}_{q^*,\alpha}$ is able to overturn the conclusion of such a test. In the rare cases when $\text{RV}_{q^*,\alpha} \leq 1 - 1/f^2_{q^*}$, setting $\text{RV}_{q^*,\alpha} = (f^2_{q^*} - f^{*2}_{\alpha,\text{df}-1})/(1 + f^2_{q^*})$ restores the property that no confounder weaker on both associations would change the conclusion.[12] Note that, since we are considering sample uncertainty, $\text{RV}_{q^*,\alpha}$ is a more conservative measure than RV_{q^*}. For a fixed $|t^*_{\alpha,\text{df}-1}|$, $\text{RV}_{q^*,\alpha}$

[12] This occurs when the variance reduction due to an increase in $R^2_{Y\sim Z|D,\boldsymbol{X}}$ dominates its effect on the bias. Such cases are unlikely in practice, see appendix 7.1.1 for details.

converges to RV_{q^*} when the sample size grows to infinity.

The $R^2_{Y\sim D|\boldsymbol{X}}$ as an extreme scenario analysis

The second measure we propose is the proportion of variation in the outcome uniquely explained by the treatment—$R^2_{Y\sim D|\boldsymbol{X}}$. Consider the following question: "if an extreme confounder explained all the residual variance of the outcome, how strongly associated with the treatment would it need to be in order to eliminate the estimated effect?" As it happens, the answer is precisely the $R^2_{Y\sim D|\boldsymbol{X}}$.

Specifically, a confounder explaining *all* residual variance of the outcome implies we have $R_{Y\sim Z|D,\boldsymbol{X}} = 1$. By Equation 2.14, to bring the estimated effect down to zero (relative bias = 1), this means $|f_{D\sim Z|\boldsymbol{X}}|$ needs to equal $|f_{Y\sim D|\boldsymbol{X}}|$, which implies $R^2_{D\sim Z|\boldsymbol{X}} = R^2_{Y\sim D|\boldsymbol{X}}$. Thus, $R^2_{Y\sim D|\boldsymbol{X}}$ is not only the determinant of the robustness of the treatment effect coefficient, but can also be interpreted as the result of an "extreme scenario" sensitivity analysis.

2.4.4 Bounding the strength of the confounder using observed covariates

Arguably, the most difficult part of a sensitivity analysis is taking the description of a confounder that would be problematic from the formal results, and reasoning about whether a confounder with such strength plausibly exists in one's study, given its design and the investigator's contextual knowledge. In this section, we introduce a novel bounding approach that can help alleviate this difficulty. The rationale for the method is the realization that, while in some cases an investigator may not be able make direct plausibility judgments about the strength of an unobserved confounder Z, she might still have grounds to make judgements about *its relative strength*, for instance, claiming that Z cannot possibly account for as much variation of the treatment assignment as some observed covariate X. How can we formally codify and leverage these claims regarding relative strength (or importance) of covariates for sensitivity analysis?

Clearly, there is not a unique way to measure the relative strength of variables [88]. For the task at hand, however, any proposal must meet the minimal criterion of solving the correct identification problem—essentially, this means the chosen measure of relative strength must be sufficient to identify (or bound) the bias, and a new function (or bound) in terms of that measure must be derived [37]. Previous work has proposed informal benchmarking procedures that fail this minimal criterion and can generate misleading sensitivity analysis results, even if researchers had correct knowledge about the relative strength of Z (58, 81, 60, 17, 50, 26, 100). We elaborate on the pitfalls of this informal approach in Section 2.6.2 of the Discussion.

Additionally, simply obtaining a formal identification result is not enough for it to be useful in applied settings—investigators must still be able to reason cogently about whether confounders are "stronger" than observed covariates using the chosen measure of relative strength. Since this depends on context, it is highly desirable to have a variety of measures for those relative comparisons (allowing researchers to choose the ones that are best suited for a given analysis) and that those measures have relevant interpretations [88]. An example of the risks entailed by ignoring this requirement can be found in the coefficient of "proportional selection on observables" advanced by [104], which will be discussed in Section 2.6.3.

With this in mind, here we offer three main alternatives to bound the strength of the unobserved confounder, by judging: (i) how the *total* R^2 of the confounder compares with the *total* R^2 of a group of observed covariates; (ii) how the *partial* R^2 of the confounder compares with the *partial* R^2 of a group of observed covariates, having taken into account the explanatory power of remaining observed covariates; or, (iii) how the *partial* R^2 of the confounder compares with the *partial* R^2 of a group of observed covariates, having taken into account the explanatory power of remaining observed covariates *and* the treatment assignment. These are natural measures of relative importance for OLS, and can be interpreted as comparisons of the consequences of dropping a (group of) variable(s) in variance reduction or prediction error [88].

The choice of bounding procedures one should use depends on which of these quantities the investigator prefers and can most soundly reason about in their own research. In our running example, within a given village, one may argue that *Female* is the most important visible characteristic that could be used for exposure to violence, and it likely explains more of the residual variation in targeting than could any unobserved confounder. For this reason (as well as simplicity of exposition) in the main text we illustrate the use of the third type of bound, but we refer readers to appendix 7.1.3 for further discussion and derivations of the other two variants.[13]

Assume $Z \perp \boldsymbol{X}$, or, equivalently, consider only the part of Z not linearly explained by $\boldsymbol{X}$. Now suppose the researcher believes she has measured the key determinants of the outcome and treatment assignment process, in the sense that the omitted variable cannot explain as much residual variance (or cannot explain a large multiple of the variance) of D or Y in comparison to a observed covariate X_j. More formally, define k_D and k_Y as,

$$k_D := \frac{R^2_{D\sim Z|\boldsymbol{X}_{-j}}}{R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}, \qquad k_Y := \frac{R^2_{Y\sim Z|\boldsymbol{X}_{-j},D}}{R^2_{Y\sim X_j|\boldsymbol{X}_{-j},D}}. \tag{2.21}$$

Where $\boldsymbol{X}_{-j}$ represents the vector of covariates $\boldsymbol{X}$ excluding X_j. That is, k_D indexes how much variance of the treatment assignment the confounder explains relative to how much X_j explains (after controlling for the remaining covariates). To make things concrete, for example, if the researcher believes the omission of X_j would result in a larger mean squared error of the treatment assignment regression than the omission of Z, this equals the claim $k_D \leq 1$. The same reasoning applies to k_Y.

Given parameters k_D and k_Y, we can rewrite the strength of the confounders as,

$$R^2_{D\sim Z|\boldsymbol{X}} = k_D f^2_{D\sim X_j|\boldsymbol{X}-j}, \qquad R^2_{Y\sim Z|D,\boldsymbol{X}} \leq \eta^2 f^2_{Y\sim X_j|\boldsymbol{X}_{-j},D} \tag{2.22}$$

where η is a scalar which depends on k_Y, k_D and $R^2_{D\sim X_j|\boldsymbol{X}-j}$, (see appendix 7.1.3 for details). These equations allow us to investigate the maximum effect a confounder at most "k times" as strong as a particular covariate X_j would have on the coefficient estimate. These results are

[13] Another reason we employ this type of bound in the main text is that it is most closely related to approaches used by other sensitivity analyses to which we contrast our results. These include the informal benchmarks of [81] as well as the bounding proposal of [104], discussed in Section 2.6.

also tight, in the sense that we can always find a confounder that makes the second inequality an equality. Further, certain values for k_D and k_Y may be ruled out by the data (for instance, if $R^2_{D\sim X_j|\boldsymbol{X}_{-j}} = 50\%$ then k_D must be less than 1).

Our bounding exercises can be extended to any subset of the covariates. For instance, the researcher can bound the effect of a confounder as strong as *all* covariates $\boldsymbol{X}$ or any subset thereof. The method can also be extended to allow different subgroups of covariates to bound $R^2_{D\sim Z|\boldsymbol{X}}$ and $R^2_{Y\sim Z|D,\boldsymbol{X}}$— thus, if a group of covariates $\boldsymbol{X}_1$ is known to be the most important driver of selection to treatment, and another group of covariates $\boldsymbol{X}_2$ is known to be the most important determinant of the outcome, the researcher can exploit this fact.

2.4.5 Sensitivity to multiple confounders

The previous results let us assess the bias caused by a single confounder. Fortunately, they also provide *upper bounds* in the case of *multiple* unobserved confounders.[14] Allowing $\boldsymbol{Z}$ to be a set (matrix) of confounders and $\hat{\boldsymbol{\gamma}}$ its coefficient vector, the full equation we wished we had estimated becomes

$$Y = \hat{\tau}D + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\gamma}} + \hat{\varepsilon}_{\text{full}}. \tag{2.23}$$

Now consider the single variable $Z^* = \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$. The bias caused by omitting $\boldsymbol{Z}$ is the same as omitting the linear combination Z^*, and one can think about the effect of multiple confounders in terms of this single confounder. Estimating the regression with $\boldsymbol{X}$ and Z^* instead of $\boldsymbol{X}$ and $\boldsymbol{Z}$ gives the same results for $\hat{\tau}$,

$$Y = \hat{\tau}D + \boldsymbol{X}\hat{\boldsymbol{\beta}} + Z^* + \hat{\varepsilon}_{\text{full}}. \tag{2.24}$$

Accordingly, Z^* has the same partial R^2 with the outcome as the full set $\boldsymbol{Z}$. However, the partial R^2 of Z^* with the treatment must be less than or equal to the partial R^2 of $\boldsymbol{Z}$ with

[14] See [76], Section 4.1, for an alternative proof.

the treatment—this follows simply because the choice of the linear combination $\hat{\boldsymbol{\gamma}}$ is the one the maximizes the R^2 with the outcome, and not with the treatment. Hence, the bias caused by a multivariate $\boldsymbol{Z}$ must be less than or equal the bias computed using Equation 2.13.

A similar reasoning can be applied to the standard errors. Since the effective partial R^2 of the linear combination Z^* with the treatment is less than that of $\boldsymbol{Z}$, simply modifying sensitivity Equation 2.12 to account for the correct degrees of freedom ($\text{df} - k$ instead of $\text{df} - 1$) will give conservative adjusted standard errors for a multivariate confounder. From a practical point of view, however, we note that further correction of the degrees of freedom might be an unnecessary formality—we are performing a hypothetical exercise, and one can always imagine to have measured Z^*.

Finally, note the set of confounders $\boldsymbol{Z}$ is arbitrary, thus it accommodates nonlinear confounders as well as misspecification of the functional form of the observed covariates $\boldsymbol{X}$. To illustrate the point, let $Y = \hat{\tau}D + \hat{\beta}X + \hat{\gamma}_1 Z + \hat{\gamma}_2 Z^2 + \hat{\gamma}_3(Z \times X) + \hat{\gamma}_4 X^2 + \hat{\varepsilon}_{\text{full}}$, and imagine the researcher did not measure Z and did not consider that X could also enter the equation with a squared term. Now just call $\boldsymbol{Z} = (Z_1 = Z, Z_2 = Z^2, Z_3 = Z \times X, Z_4 = X^2)$ and all the previous arguments follow.

2.5 Using the partial R^2 parameterization for sensitivity analysis

Returning to our running example of violence in Darfur, we illustrate how these tools can be deployed in an effort to answer the following questions: (i) How strong would a particular confounder (or group of confounders) have to be to change our conclusions? (ii) In a worst case scenario, how vulnerable is our result to *many* or *all* unobserved confounders acting together, possibly non-linearly? (iii) Are the confounders that would alter our conclusions plausible, or at least how strong would they have to be relative to observed covariates?

2.5.1 Proposed minimal reporting

Table 2.1 illustrates the type of reporting we propose should accompany linear regression models used for causal inference with observational data. Along with traditionally reported statistics (point estimate, standard error and t-value), we propose researchers present (i) the partial R^2 of the treatment with the outcome, and (ii) the robustness value, RV, both for where the point estimate and the confidence interval would cross zero (or another meaningful reference value).[15] Finally, in order to aid user judgment, we encourage researchers to provide plausible bounds on the strength of the confounder. These may be based upon bounds employing meaningful covariates determined by the research context and design (Section 2.4.4), or in principle may be available from theory and previous literature.

Outcome: *Peace Index*

Treatment:	Estimate	Std. Error	t-value	$R^2_{Y\sim D\mid\boldsymbol{X}}$	RV	$\text{RV}_{\alpha=0.05}$
Directly Harmed	0.097	0.023	4.18	2.2%	13.9%	7.6%

df = 783, Bound (Z as strong as *Female*): $R^2_{Y\sim Z\mid D,\boldsymbol{X}} = 12\%$, $R^2_{D\sim Z\mid\boldsymbol{X}} = 1\%$

Table 2.1: Proposed minimal reporting on sensitivity to unobserved confounders.

For our running example of violence in Darfur, Table 2.1 shows an augmented regression table, including the robustness value (RV) of the *Directly Harmed* coefficient, 13.9%. This means that unobserved confounders explaining at least 13.9% of the residual variance of both the treatment and the outcome would explain away the estimated treatment effect. It also means that any confounder explaining less than 13.9% of the residual variance of both the treatment and the outcome would not be strong enough to bring down the estimated effect to zero. For cases where one association is over 13.9% and the other is below, we conduct additional analyses illustrated in the next subsection. Nevertheless, the RV still fully characterizes the robustness of the regression coefficient to unobserved confounding—it provides a quick, meaningful reference point for understanding the minimal strength of bias necessary to overturn the research conclusions.[16]

[15] For convenience, we refer to the RV_{q^*} or $\text{RV}_{q^*,\alpha}$ with $q^* = 1$ as simply the RV or RV_α

[16] That is, any confounder with an equivalent bias factor of $\text{BF} = \text{RV}/\sqrt{1-\text{RV}}$.

Adjusting for confounding may not bring the estimate to zero, but rather into a range where it is no longer "statistically significant." Therefore, the robustness value accounting for statistical significance, $\text{RV}_{\alpha=0.05}$, is also shown in the table. For a significance level of 5%, the robustness value goes down from 13.9% to 7.6%—that is, confounders would need to be only about half as strong to make the estimate not "statistically significant." Finally, the partial R^2 of the treatment with the outcome, $R^2_{Y\sim D|\boldsymbol{X}}$, in Table 2.1 gives a sensitivity analysis for an extreme scenario: if confounders explained 100% of the residual variance of the outcome, they would need to explain at least 2.2% of the residual variance of the treatment to bring down the estimated effect to zero.

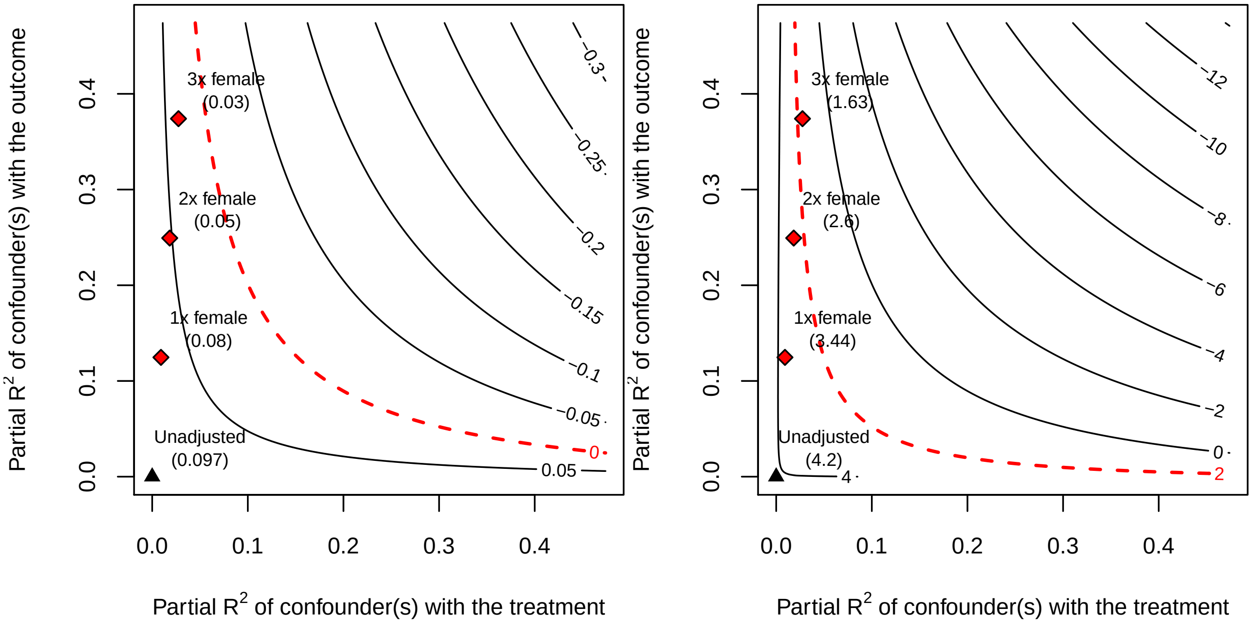
Confronted with those results, we now need to judge whether confounders with the strengths revealed to be problematic are plausible. If one can claim to have measured the most important covariates in explaining treatment and outcome variation, it is possible to bound the strength of the confounder with the tools of Section 2.4.4 and judge where it falls relative to these quantities. The lower right corner of Table 2.1 shows the strength of association that a confounder as strong as *Female* would have: $R^2_{Y\sim Z|D,\boldsymbol{X}} = 12\%$ and $R^2_{D\sim Z|\boldsymbol{X}} = 1\%$. As the robustness value is higher than either quantity, the table readily reveals that such a confounder could not fully eliminate the point estimate. In addition, since the bound for $R^2_{D\sim Z|\boldsymbol{X}}$ is less than $R^2_{Y\sim D|\boldsymbol{X}} = 2.2\%$, a "worst case confounder" explaining *all* of the left-out variance of the outcome and as strongly associated with the treatment as *Female* would not eliminate the estimated effect either.

Domain knowledge about how the treatment was assigned or regarding the main determinants of the outcome is required to make any such comparisons meaningful. In our running example, a reasonable argument can be made that gender is one of the most visually apparent characteristic of an individual during the attacks, and that, within village, gender was potentially the most important factor to explain targeting due to the high level of sexual violence. Thus, if one can argue that total confounding as strongly associated with the treatment as *Female* is implausible, those bounding results show it cannot completely account for the observed estimated effect.

These sensitivity exercises are exact when considering a single linear unobserved confounder and are conservative for multiple unobserved confounders, possibly acting non-linearly—this includes the explanatory power of *all left out factors*, even misspecification of the functional form of observed covariates. It is worth pointing out that sensitivity to any arbitrary confounder with a given pair of partial R^2 values $(R^2_{Y\sim Z|D,\boldsymbol{X}}, R^2_{D\sim Z|\boldsymbol{X}})$ can also be easily computed with the information on the table.

2.5.2 Sensitivity contour plots with partial R^2: estimates and t-values

The next step is to refine the analysis with tools that visually demonstrate how confounders of different types would affect point estimates and t-values, while showing where bounds on such confounders would fall under different assumptions on how unobserved confounders compare to observables.[17]



(a) Sensitivity contour plot of the point estimate.

(b) Sensitivity contour plot of the t-value.

Figure 2.3: Sensitivity contour plots in the partial R^2 scale with benchmark bounds.

[17] Here we focus on the plots for point estimates and t-values, but note p-values can be obtained from the t-values, and the confidence interval end-points by adjusting the estimate with the appropriate multiple of the standard-errors.

Perhaps the first plot investigators would examine would be one similar to Figure 2.2, but now in the partial R^2 parameterization (Figure 2.3a). The horizontal axis describes the fraction of the residual variation in the treatment (partial R^2) explained by the confounder; the vertical axis describes the fraction of the residual variation in the outcome explained by the confounder.[18] The contours show the adjusted estimate that would be obtained for an unobserved confounder (in the full model) with the hypothesized values of the sensitivity parameters (assuming the direction of the effects hurts our preferred hypothesis).

While the contour plot used in illustrating the traditional OVB approach focused on a specific binary confounder—*Center*—the contour plot with the partial R^2 parameterization allows us to assess sensitivity to any confounder, irrespective of its unit of measure. Additionally, since the sensitivity equations give an upper bound for the multivariate case, the same plot can be used to assess the sensitivity to any *group* of confounders, here including non-linear terms, such as the example of *Political Attitudes* and *Wealth* acting together. Notice that if we choose a contour of interest (such as where the effect equals zero), and find the point with equal values on the horizontal and vertical axes (i.e. where it crosses a 45-degree line), this correspond to the robustness value. That is, the RV is a convenient, interpretable summary of a critical line of the contour plot.

Further, the bounding exercise results in points on the plot showing the bounds on the partial R^2 of the unobserved confounder if it were k times "as strong" as the observed covariate *Female*. The first point shows the bounds for a confounder (or group of confounders) as strong as *Female*, as was also shown in Table 2.1. A second reference point shows the bounds for confounders *twice* as strong as *Female*, and finally the last point bounds the strength of confounders *three times* as strong as *Female*. The plot reveals that the *sign of the point estimate* is still relatively robust to confounding with such strengths, although the magnitude would be reduced to 77%, 55% and 32% of the original estimate, respectively.

[18] As discussed in Section 2.4.2, axes could be transformed to show instead (i) the total R^2 including the confounders $R^2_{Y\sim D+\boldsymbol{X}+Z}$ and $R^2_{D\sim\boldsymbol{X}+Z}$, (ii) the difference in the total R^2 including the confounders, i.e., $R^2_{Y\sim D+\boldsymbol{X}+Z} - R^2_{Y\sim D+\boldsymbol{X}}$ and $R^2_{D\sim\boldsymbol{X}+Z} - R^2_{D\sim\boldsymbol{X}}$, (iii) the partial correlations (by simply taking the square-root), (iv) the partial R^2 of the confounder with the outcome *not* conditioning on the treatment, among other options that may aid interpretation.

Moving to inferential concerns, Figure 2.3b now shows the sensitivity of the *t-value* of the treatment effect. As we move along the horizontal axis, not only the adjusted effect reduces, but we also get larger standard-errors due to the variance inflation factor of the confounder. If we take the t-value of 2 as our reference (the usual approximate value for a 95% confidence interval), the plot reveals the statistical significance of *Directly Harmed* is robust to a confounder as strong as, or twice as strong as *Female*. However, whereas confounders *three times* as strong as *Female* would not erode the point estimate to zero, we cannot guarantee the estimate would remain "statistically significant" at the 5% level.

Altogether, these bounding exercises naturally lead to the questions: are such confounders plausible? Do we think it possible that confounders might exist that are three times as strong as *Female*? If so, what are they? While one may not have complete confidence in answering such questions, we have moved the discussion from a qualitative argument about whether any confounding is possible to a more disciplined, quantitative argument that entices researchers to think about possible threats to their research design.

2.5.3 Sensitivity plots of extreme scenarios

Even with a good understanding of the treatment assignment mechanism, investigators may not always be equipped to convincingly limit the association of the confounder with the outcome. In such cases, exploring sensitivity analysis to extreme-scenarios is still an option. If we set $R^2_{Y\sim Z|D,\boldsymbol{X}}$ to one or some other conservative value, how strongly would such a confounder need to be associated with the treatment in order to problematically change our estimate? While in some cases this exercise could reveal that confounders weakly related to the treatment would be sufficient to overturn the estimated effect, survival to extreme scenarios may help investigators demonstrate the robustness of their results.

Applying this to our running example, results are shown in Figure 2.4. The solid curve represents the case where unobserved confounder(s) *explain all the left-out residual variance of the outcome*. On the vertical axis we have the adjusted treatment effect, starting from the case with no bias and going down as the bias increases, reducing the estimate; the horizontal

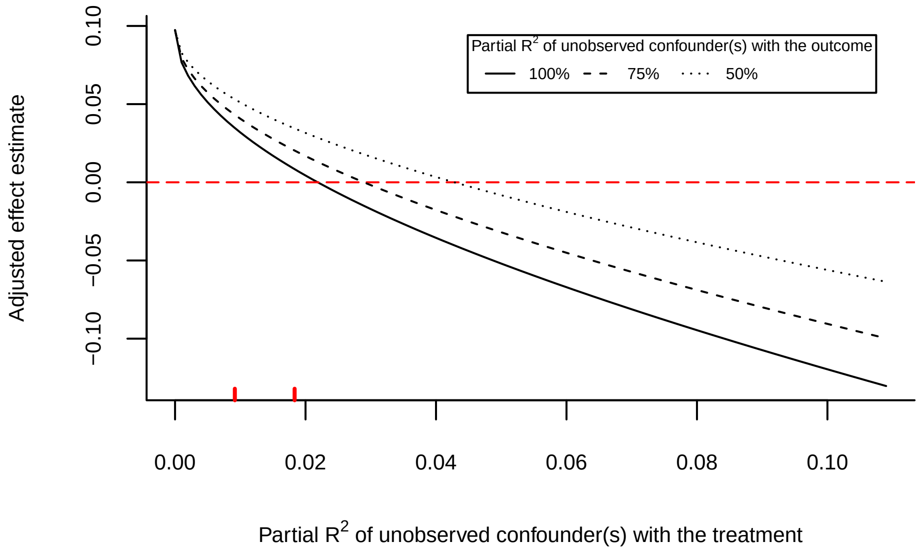


Figure 2.4: Sensitivity analysis to extreme scenarios.

axis shows the partial R^2 of the confounder with the treatment. In this *extreme scenario*, as we have seen, $R^2_{D\sim Z|\boldsymbol{X}}$ would need to be exactly the same as the partial R^2 of the treatment with the outcome to bring down the estimated effect to zero—that is, it would need to be at least 2.2%, a value below the bound for a confounder once or twice as strong as *Female* (shown by red tick marks), which in this case is arguably one of the strongest predictors of the treatment assignment. In most circumstances, considering the worst case scenario of $R^2_{Y\sim Z|D,\boldsymbol{X}} = 1$ might be needlessly conservative. Hence, we propose plotting other extreme scenarios, as shown in Figure 2.4, where we consider different values of the partial R^2 of the unobserved confounder with the outcome, including 75% and 50%.

2.6 Discussion

2.6.1 Making formal sensitivity analysis standard practice

Given that ruling out unobserved confounders is often difficult or impossible in observational research, one might expect that sensitivity analyses would be a routine procedure in numerous disciplines. Why then are they not commonplace? We surmise there are three main obstacles, which we directly address in this chapter.

Strong parametric assumptions

First, the assumptions that many methods impose on the nature and distribution of unobserved confounders as well as on the treatment assignment mechanism may be difficult to sustain in some cases. For instance, [124], [81], [26] and [50] require specifying the distribution of the confounder as well as modeling the treatment assignment mechanism; in another direction, the methods put forward in [118], [21], [17] need to directly specify a confounding function parameterizing the difference in potential outcomes among treated and control units. While assessing the sensitivity to some forms of confounding is an improvement over simply assuming no confounding (and users may be able to make suitable parameteric assumptions in some circumstances), widespread adoption of sensitivity analysis would benefit from methods that do not require users to make those restrictions *a priori*. Our derivations are rooted in the traditional OVB precisely to avoid those simplifying assumptions. As we have seen, the partial R^2 parameterization allows a flexible framework for assessing the sensitivity of the point estimate, as well as t-values and confidence intervals, allowing for multiple (possibly nonlinear) confounders, even including mispecification of the functional form of the observed covariates.

Lack of simple sensitivity measures for routine reporting

A second obstacle to a wider adoption of sensitivity analysis is the lack of general, yet simple and interpretable sensitivity measures users can report alongside other regression summary statistics. Our minimal reporting recommendation for regression tables (see Table 2.1) aims to fill this gap for regression models with: (i) the robustness value, which conveniently summarizes the minimal strength of association a confounder needs to have to change the research conclusions, and (ii) the $R^2_{Y\sim D|\boldsymbol{X}}$, which works as an extreme-scenario sensitivity analysis. Regarding the robustness value in particular, we now discuss its relation to two other proposals advocated in the literature: the *impact thresholds* of [58] and the E-value of [142].

Frank [58] proposes characterizing the strength of the unobserved confounder Z with what he denotes as its *impact*, defined as the product $R_{Y\sim Z|\boldsymbol{X}} \times R_{D\sim Z|\boldsymbol{X}}$.[19] This is then used to determine *impact thresholds*, defined as the minimum impact of the unobserved confounder necessary to not reject the null hypothesis of *zero effect*. However, as Equation 2.14 reveals, the determinant of the bias is the bias factor $\text{BF} = R_{Y\sim Z|D,\boldsymbol{X}} \times f_{D\sim Z|\boldsymbol{X}}$, which does not have a one-to-one mapping with the confounder's impact. This can be made clear by rewriting the relative bias showing the product $R_{Y\sim Z|\boldsymbol{X}} \times R_{D\sim Z|\boldsymbol{X}}$ explicitly,

$$\text{relative bias} = \frac{|\overbrace{R_{Y\sim Z|\boldsymbol{X}} R_{D\sim Z|\boldsymbol{X}}}^{\text{Frank's } \textit{impact}} - R_{Y\sim D|\boldsymbol{X}} R^2_{D\sim Z|\boldsymbol{X}}|}{|R_{Y\sim D|\boldsymbol{X}}(1 - R^2_{D\sim Z\boldsymbol{X}})|}. \tag{2.25}$$

Equation 2.25 reveals that: (i) an unobserved confounder with *zero impact* can still cause non-zero (downward) bias; (ii) an unobserved confounder with a *non-zero impact* can nevertheless induce zero bias (when impact $= R_{Y\sim D|\boldsymbol{X}} R^2_{D\sim Z|\boldsymbol{X}}$); and, (iii) the two terms that compose the product $R_{Y\sim Z|\boldsymbol{X}} \times R_{D\sim Z|\boldsymbol{X}}$ do not enter symmetrically in the bias equation, hence confounders with the *same impact* can cause *widely different biases*. This creates difficulties when trying to generalize the impact thresholds proposed in [58] to arbitrary non-zero null hypothesis of regression coefficients.[20] Note this is not a problem for the robustness value, since it acts as a convenient reference point uniquely characterizing any confounder with a bias factor of $\text{BF} = \text{RV}_{q^*}/\sqrt{1 - \text{RV}_{q^*}}$.

As to [142], the authors have recently advanced the E-value, a sensitivity measure suited specifically for the *risk ratio*. For other effect measures, such as risk differences, the E-value is an approximation, whereas if the researcher uses linear regression to obtain an estimate, the robustness value is exact. Also, while the robustness value parameterizes the association of the confounder with the treatment and the outcome in terms of percentage of variance

[19] Not to confuse with $\hat{\gamma}$ of the "impact times imbalance" heuristic, as discussed in Section 2.3.2.

[20] Let q^* denote the relative bias of interest and consider biases that move the effect toward (or through) zero. Solving Equation 2.25 for *impact* gives us impact $= R_{Y\sim D|\boldsymbol{X}}(q^* - (q^* - 1)R^2_{D\sim Z|\boldsymbol{X}})$. Note that, given q^* and $R_{Y\sim D|\boldsymbol{X}}$, the *impact* necessary to bring about a relative bias of magnitude q^* still depends on the sensitivity parameter $R^2_{D\sim Z|\boldsymbol{X}}$—except when $q^* = 1$. For a numerical example, see appendix 7.1.2.1.

explained (the partial R^2), the E-value parameterizes these in terms of risk ratios. Whether one scale is preferable over the other depends on context, and researchers should be aware of both options. Overall, we believe the dissemination of measures such as the E-value and the robustness value is an important step towards the widespread adoption of sensitivity analysis to unobserved confounding. In current practice, robustness is often informally or implicitly linked to t-values or p-values, neither of which correctly characterizes how sensitive an estimate is to unobserved confounding. The extension of the robustness value to non-linear models is worth exploring in future research.

Difficulty in connecting sensitivity analysis to domain knowledge

Finally, the third and perhaps most fundamental obstacle to the use of sensitivity analysis is the difficulty in connecting the formal results to the researcher's substantive understanding about the object under study. This can be only partially overcome by statistical tools, as it relies upon the nature of the domain knowledge used for plausibility judgments. In this chapter we have showed how one can formally bound the strength of an unobserved confounder with the same strength (or a multiple thereof) as a chosen group of observed covariates, using three different types of comparisons. This allows researchers to exploit knowledge regarding the relative importance of observed covariates: when researchers can credibly argue to have measured the most important determinants of the treatment assignment and of the outcome (in terms of variance explained), this bounding exercise can be a valuable tool. As we discuss next, previous attempts to make such comparisons have been problematic, either due to informal benchmarking practices that do not warrant the claims they purport to make, or by relying on inappropriate parameterization choices.

2.6.2 The risks of informal benchmarking

While prior work has suggested informal benchmarking procedures using statistics of observed covariates $\boldsymbol{X}$ to help researchers "calibrate" their intuitions about the strength of the unobserved confounder Z (58, 81, 76, 50, 26, 100, 75), this practice has undesirable properties and can lead users to erroneous conclusions, even in the ideal case where they do have the correct knowledge about how Z compares to $\boldsymbol{X}$. This happens because the estimates of how the observed covariates are related to the outcome may be themselves affected by the omission of Z, regardless of whether one assumes Z to be independent of $\boldsymbol{X}$. To illustrate this threat concretely, let us first consider a simple simulation where there is no effect of D on Y, Z is orthogonal to X and, more importantly, Z *is exactly like* X.[21] The results are shown in Figure 2.5.

Note the informal benchmark point is still far away from zero, leading the investigator to incorrectly conclude that a confounder "not unlike X" would not be sufficient to bring down the estimated effect to zero—when in fact it would. This incorrect conclusion occurs *despite* the investigator *correctly assuming* both that the unobserved confounder is "no worse" than X (in terms of its strength of relationship to the treatment and outcome) and that $Z \perp X$. Figure 2.5 also shows the formal bounds obtained with the procedures given in Section 2.4.4. Note these would lead the researcher to the correct conclusion: an unobserved confounder with the same strength as X would be powerful enough to bring down the estimate to zero.

Why exactly does this happen? Consider for a moment the difference between the coefficient on $\boldsymbol{X}$ in the full Equation 2.3, $\hat{\boldsymbol{\beta}}$, and its estimate in the restricted Equation 2.4, $\hat{\boldsymbol{\beta}}_{\text{res}}$. Using the same OVB approach of "impact times imbalance", we arrive at $\hat{\boldsymbol{\beta}}_{\text{res}} - \hat{\boldsymbol{\beta}} = \hat{\gamma}\hat{\boldsymbol{\psi}}$, where $\hat{\boldsymbol{\psi}}$ is obtained from the regression $Z = \hat{\delta} D + \boldsymbol{X}\hat{\boldsymbol{\psi}} + \hat{\varepsilon}_Z$. Note that $\hat{\boldsymbol{\psi}}$ can be non-zero even if $\boldsymbol{X} \perp Z$, because D is a collider [109], and conditioning on D creates dependency between Z and $\boldsymbol{X}$. The reasoning holds whether one is using the regression coefficients themselves or other observed statistics, such as partial correlations, partial R^2 values or t-values. This renders claims of the type "a confounder Z not unlike X could not change the research conclusions" unreliable when observed statistics without proper adjustment are used for benchmarking.

We can use the formal bounds derived in Equation 2.22 to quantify how misleading claims

[21] We use structural equations, $Y = X + Z + \varepsilon_y$, $D = X + Z + \varepsilon_d$, $X = \varepsilon_x$, $Z = \varepsilon_z$ where all disturbances, are independent standard normal random variables. See also appendix 7.1.4.

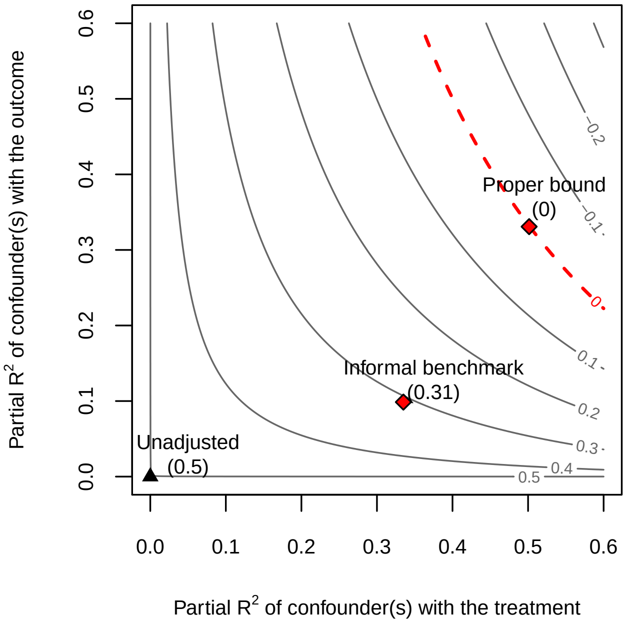


Figure 2.5: Sensitivity contours, point estimate. Informal benchmarking *versus* proper bound.

using informal benchmarks would be. In the partial R^2 parameterization, this amounts to using as benchmarks $k_D R^2_{D\sim X_j|\boldsymbol{X}-j}$ and $k_Y R^2_{Y\sim X_j|\boldsymbol{X}_{-j},D}$, instead of the proper bounds $k_D f^2_{D\sim X_j|\boldsymbol{X}-j}$ and $\eta^2 f^2_{Y\sim X_j|\boldsymbol{X}_{-j},D}$. There are, thus, two discrepancies: (i) an adjustment of baseline variance to be explained, when converting the partial R^2 to partial Cohen's $f^2 = R^2/(1-R^2)$, which affects both coordinates of the benchmark; and, (ii) the collider bias due to the association of X_j with D, which affects only the bound on $R^2_{Y\sim Z|D,\boldsymbol{X}}$ via $\eta^2 \geq k_Y$.[22] Therefore, the stronger the association of X_j with the treatment, and the larger the multiples used for comparisons ("k times as strong"), the more misleading informal benchmarks will be.[23] We thus advise against informal benchmarking procedures, and previous studies relying upon these methods may warrant revisiting, especially those where benchmark points have strong association with the treatment assignment.

[22] The adjustment of baseline variance may affect informal benchmarks based on correlational [58], partial R^2 [81], and t-value [76] measures. The collider bias may affect informal benchmarks that condition on D. Benchmarks that do not condition on D (such as in 58) are not affected by collider bias.

[23] In our running example, since *female* explains less than 1% of the residual variance of the treatment, informal benchmarks would not be markedly different from the formal ones.

2.6.3 On the choice of parameterization

The approach of [76] is also rooted in the OVB framework, but it suffers from two main deficiencies. The first is the central role informal benchmarking plays in their proposal, which can be seriously misleading as discussed in the previous section. The second issue is more subtle, but equally important: the choice of parameterization. [76] ask researchers to "calibrate intuitions" about the strength of the confounder with the treatment using a t-value. This is a problematic choice because the t-value incorporates information on both the strength of association and the sample size, the latter being irrelevant for identification concerns. What constitutes a large t-value for "statistical significance" does not map directly to what constitutes a large strength of a confounder, as this mapping varies significantly depending on sample size.[24]

An alternative bounding argument has also been presented in [104] which, unlike the informal benchmarking practices previously discussed, provides a formal identification result. Nevertheless, the proposed procedure asks users to reason about a quantity that is very difficult to understand. More precisely, [104] asks researchers to make plausibility judgments on two sensitivity parameters, $R_{\max}$ and δ_{Oster}. The $R_{\max}$ parameter is simply the maximum explanatory power that one could have with the full outcome regression, i.e., $R_{\max} = R^2_{Y\sim D+\boldsymbol{X}+\boldsymbol{Z}}$. As discussed in Section 2.4.2 (Equation 2.17) this has a one to one relationship with $R^2_{Y\sim \boldsymbol{Z}|\boldsymbol{X},D}$,

$$R^2_{Y\sim \boldsymbol{Z}|\boldsymbol{X},D} = \frac{R_{\max} - R^2_{Y\sim D+\boldsymbol{X}}}{1 - R^2_{Y\sim D+\boldsymbol{X}}} \tag{2.26}$$

By contrast the second sensitivity parameter, δ_{Oster}, is not easily interpretable in sub-

[24] The t-value in the expression of the bias is an artifact of both multiplying and dividing by the degrees of freedom, as in our Equation 2.12. While t-values can be useful for computational purposes (to utilize quantities routinely reported in regression tables), their dependence on sample size makes them inappropriate for contemplating how strongly related a confounder is to the treatment. Consider a t-value of 200. With 100 degrees of freedom, the confounder explains virtually all the residual variance of the treatment (partial R^2 of 0.9975), while with 10 million degrees of freedom, the confounder explains less than 0.5%. These are clearly confounders with very different strengths, and the partial R^2 clarifies this distinction.

stantive terms. Following [1], [104] defines "indices" $W_1 := \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $W_2 := \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$, where $\boldsymbol{X}$ is a matrix of observed covariates and $\boldsymbol{Z}$ a matrix of unobserved covariates. Critically, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are chosen such that $Y = \hat{\tau}D + W_1 + W_2 + \hat{\varepsilon}_{\text{full}}$.[25] The δ_{Oster} parameter equals $\text{cov}(W_2, D)/\text{var}(W2) \times \text{var}(W_1)/\text{cov}(W_1, D)$, and is intended as a measure of "proportional selection", i.e. how strongly the unobservables drive treatment assignment, relative to the observables. The problem here is that constructing indices W_1 and W_2 based on relationships to the outcome is not innocuous: δ_{Oster} captures not only the relative influence of $\boldsymbol{X}$ and $\boldsymbol{Z}$ over the treatment, but also their association with the outcome. To examine the simple case with only one covariate and one confounder and assuming $X \perp Z$, we have,

$$\delta_{\text{Oster}} = \frac{\text{cov}(W_2, D)}{\text{var}(W_2)}\frac{\text{var}(W_1)}{\text{cov}(W_1, D)} = \frac{\text{cov}(\hat{\gamma}Z, D)}{\text{var}(\hat{\gamma}Z)}\frac{\text{var}(\hat{\beta}X)}{\text{cov}(\hat{\beta}X, D)} = \frac{\text{cov}(Z, D)}{\hat{\gamma}\text{var}(Z)}\frac{\hat{\beta}\text{var}(X)}{\text{cov}(X, D)} = \frac{\hat{\lambda}}{\hat{\gamma}}\frac{\hat{\beta}}{\hat{\theta}}, \tag{2.27}$$

where $\hat{\lambda}$ and $\hat{\theta}$ are the coefficients of the regression, $D = \hat{\theta}X + \hat{\lambda}Z + \hat{\varepsilon}_D$. Consequently, claims that $\delta_{\text{Oster}} = 1$ implies "the unobservable and observables are equally related to the treatment" [104, p.6] can lead researchers astray, as this quantity also depends upon associations with the outcome. To see how, let the variables be standardized to mean zero and unit variance, and pick $\hat{\beta} = \hat{\theta} = p$, $\hat{\gamma} = \hat{\lambda} = p/2$, and $\hat{\tau} = 0$. In this case, the confounder Z has either half or one fourth of the explanatory power of X (as measured by standardized coefficients or variance explained), yet $\delta_{\text{Oster}} = 1$. While researchers may be able to make arguments about relative explanatory power of observables and unobservables in the treatment assignment process, the δ_{Oster} parameter does not correspond directly to such claims.[26] By contrast, the

[25] [104] uses population values. Here we use sample values to maintain consistency with the rest of the chapter, but this has no consequence for the argument in question.

[26] Indeed, arguments made by researchers applying [104] suggest they believe they are comparing the explanatory power of observables and unobservables over treatment assignment in terms such as correlation or variance explained, e.g. "Following the approach suggested by Altonji, Elder, and Taber (2005) and Oster (2017), we estimate that unobservable country-level characteristics would need to be 1.44 times more correlated with treatment than observed covariates to fully explain the apparent impact of grammatical gender on the level of female labor force participation; unobserved factors would need to be 3.23 times more closely linked to treatment to explain the impact of grammatical gender on the gender gap in labor force participation." [84, p.4]

parameter k_D we introduce in our bounding procedure (Section 2.4.4) captures precisely this notion of the relative explanatory power of the unobservable and observable over treatment assignment, in terms of partial R^2 or total R^2, depending on the investigator's preference.

Such parameterization choices are more than notional when they drive a wedge between what investigators can argue about and the values of the parameters these arguments imply. It is thus important that the sensitivity parameters used in these exercises be as transparent as possible and match investigators' conception of what the parameters imply. Hence, we employ R^2 based parameters, rather than t-values or quantities relating indices. The resulting sensitivity parameters not only correspond more directly to what investigators can articulate and reason about, but also lead to the rich set of sensitivity exercises we have discussed. Of course, further improvements may be possible and future research should investigate whether such flexibility can be achieved with yet more meaningful parameterizations.

The tools we propose here, like any other, have potential for abuse. We thus end with important caveats, in particular emphasizing that sensitivity analysis should not be used for automatic judgment, but as an instrument for disciplined arguments about confounding.

2.6.4 Sensitivity analysis as principled argument

Sensitivity analyses tell us what we would have to be prepared to believe in order to accept the substantive claims initially made [121, 122, 123]. The sensitivity exercises proposed here tell the researcher how strong unobserved confounding would have to be in order to meaningfully change the treatment effect estimate beyond some level we are interested in, and employ observed covariates to argue for bounds on unobserved confounding where possible. Whether we can rule out the confounders shown to be problematic depends on expert judgment. As a consequence, the research design, identification strategy as well as the story explaining the quality of the covariates used for benchmarking all play vital roles.

For this reason, we do not propose any arbitrary thresholds for deeming sensitivity statistics, such as the robustness value or the partial R^2 of the treatment with the outcome, sufficiently large to escape confounding concerns. In our view, no meaningful universal thresholds of the sort is possible to establish. In a poorly controlled regression on observational data, with no clear understanding of what (unobservables) might influence treatment uptake, it would be difficult to credibly claim that a robustness value of 15% is “good news”, since the investigator does not have the necessary domain knowledge to rule out the strength of unobserved confounders down to this level. On the other hand, in a quasi-experiment where the researcher knows the treatment was assigned in such a way that observed covariates account for almost any possible selection, a more credible case may be made that the types of confounders that would substantially alter the research conclusions are unlikely.

Similarly, we strongly warn against blindly employing covariates for bounding the strength of confounders, without the ability to argue that they are likely to be among the strongest predictors of the outcome or treatment assignment. A particular moral hazard is that weak covariates can make the apparent bounds look better. It is thus imperative for readers and reviewers to demand that researchers properly justify and interpret their sensitivity results, after which such claims can be properly debated. Sensitivity analysis is best suited as a tool for disciplined quantitative arguments about confounding, not for obviating scientific discussions by following automatic procedures.

This transition from a qualitative to a quantitative discussion about unobserved confounding can often be enlightening. As put by [123, p. 171], it may “provide grounds for caution that are not rooted in timidity, or grounds for boldness that are not rooted in arrogance.” A sensitivity analysis raises the bar for the skeptic of a causal estimate—not just any criticism is able to invalidate the research conclusions. The hypothesized unobserved confounder now has to meet certain standards of strength; otherwise, it cannot logically account for all the observed association. Likewise, it also raises the bar for defending a causal interpretation of an estimate—proponents must articulate how confounders with certain strengths can be ruled out.

A final point of concern is the potential misuse of sensitivity analysis in the gatekeeping of publications. Sensitivity analysis should not be misappropriated as a tool for inhibiting “imperfectly identified” research on relevant topics. Studies on important questions using state-of-the-art research design, which turn out to not be robust to reasonable sources of confounding, should not be dismissed. On the contrary, with sensitivity analyses, we can conduct imperfect investigations, while transparently revealing how susceptible our results are to unobserved confounders. This gives future researchers a starting point and roadmap for improving upon the robustness of these answers in their following inquiries.

CHAPTER 3

An Omitted Variable Bias Framework for Sensitivity Analysis of Instrumental Variables

3.1 Introduction

Unobserved confounding often complicates efforts to make causal claims from observational data (109, 83, 123). Instrumental variable (IV) regression offers a powerful and widely used tool to address unobserved confounding, by exploiting "exogenous" sources of variation of the treatment [145, 19, 4, 6]; IV methods have also become a vital tool in the analysis of randomized experiments with imperfect compliance [117, 12, 13, 4]. These qualities have made instrumental variables "a central part of the econometrics canon since the first half of the twentieth century" [82, p.324]. Beyond economics, instrumental variables are prominent tools in the arsenal of investigators seeking to make causal claims across the social sciences, epidemiology, medicine, genetics, and other fields (see e.g. 74, 46, 10, 22).

Yet, IV methods carry their own set of demanding assumptions. Principally, conditionally on certain observed covariates, an instrumental variable must not itself be confounded with the outcome, and it should influence the outcome *only* by influencing uptake of the treatment. These assumptions can be violated by omitted confounders of the instrument-outcome association, and by omitted "side-effects" of the instrument, which then influence the outcome through channels other than through the treatment.[1] Although in certain cases the

[1] In the recent IV literature, the first assumption is usually called *exogeneity*, *ignorability*, *unconfoundedness* or *independence* of the instrument, whereas the second assumption is called the *exclusion* restriction [6, 109, 83, 135]. In earlier econometric works, these two assumptions were often combined into one, also labeled the "exclusion restriction" [82].

IV assumptions may entail testable implications [107, 135, 87], they are often unverifiable and must be defended by appealing to domain knowledge and theoretical arguments. Whether a given IV study identifies the causal effect of interest, then, turns on debates as to whether these assumptions hold.

Particularly in recent years, economists and other scholars have adopted a more skeptical posture towards IV methods, emphasizing the importance of both defending the credibility of these assumptions as well as assessing the consequences of its failures (see e.g., 45, 73). For instance, recent extensive reviews of many instrumental variables widely-used in applied work, such as weather, religion, sibling structure or ethnolinguistic fractionalization, have cataloged several plausible violations of the exclusion restriction for such instruments [65, 99]. More worrisome, if the IV assumptions fail to hold, it is well known that the bias of the IV estimate may be *worse* than the original confounding bias of the simple regression estimate that the IV was supposed to address [18]. Therefore, researchers are also advised to perform *sensitivity analyses* to assess the degree of violation of the IV assumptions that would be required to alter the conclusions of an IV study. Although a variety of sensitivity methods for IV have been proposed [49, 1, 129, 131, 41, 143, 85, 37], such sensitivity analyses are still rare in practice.

In this chapter, we develop an omitted variable bias (OVB) framework for assessing the sensitivity of IV estimates against violations of its underlying assumptions.[2] Building on the results of Chapter 2, we develop a suite of sensitivity analysis tools for IV that: (i) has correct test size (or confidence interval coverage) regardless of instrument strength; (ii) naturally handles violations due to multiple "side-effects" and "confounders;" (iii) exploits

[2] We focus on the "just-identified" case with one treatment and one instrument. One reason for our focus is that a thorough consideration of the identification assumptions and how they may be violated is already complicated enough with a single instrument [6]. Second, and relatedly, in most applied settings, the single-instrument and single-treatment setup is the most common. For example, in a broad review of papers in the *American Economic Review* and 15 other journals of the *American Economic Association*, Young (2018) finds that 80% of IV regressions were of this type. Finally, in many "multiple instrument" studies, it is not uncommon for researchers to also report and give special focus on the analysis of their "best" instrument [6], or to combine multiple instruments into a single instrument, such as, for example, constructing an allele score in Mendelian Randomization [22]. Extension of the tools we develop here to the scenario with multiple instruments and treatments is object of future investigations.

expert knowledge to bound sensitivity parameters; and, (iv) can be easily implemented with standard software.

We first introduce two sensitivity statistics for IV estimates: (i) the *robustness value* describes the minimum strength of association (in terms of partial R^2) that omitted variables (side-effects or confounders) need to have, both with the instrument and with the untreated potential outcome, such that they are capable of changing the conclusions of the study; and (ii) the *extreme robustness value*, which describes the minimal strength of association that omitted variables need to have with the *instrument alone* (regardless of their association with the untreated potential outcome) in order to be problematic. We propose the routine reporting of those quantities to improve the transparency and facilitate the assessment of the credibility of IV studies. Next, we offer intuitive graphical tools for investigators to assess how postulated confounding of any degree would alter the IV hypothesis tests, as well as lower or upper limits of confidence intervals. Finally, these tools can be supplemented with formal bounds on the worst possible bias that side-effects or confounders could cause, under the assumption that the maximum explanatory power of these omitted variables are no stronger than a multiple of the explanatory power of one or more observed variables.

Conveniently, considering that investigators are already advised to carefully examine their "first stage" (the effect of the instrument on the treatment) and "reduced form" (the effect of the instrument on the outcome) (e.g. 5, 6), we show that many pivotal conclusions regarding the sensitivity of the IV estimate can in fact be reached simply through separate sensitivity analyses of these two familiar auxiliary OLS estimates. First, if researchers are interested in the null hypothesis of *zero effect,* all the OVB tools developed developed in Chapter 2 can simply be directly applied to the reduced-form regression, and confounders or side-effects shown to be problematic there are equally problematic for IV. Second, if interest lies in assessing not just the null of zero, but biases that bring the estimate partway to zero or beyond it, then the robustness of the IV estimate formally reduces to the minimum of the robustness of the reduced-form and the robustness of the first-stage regressions.

A final contribution of this chapter is that, while developing OVB tools for IV, we extended the previous OVB results for OLS providing a new way to perform sensitivity analysis that simply replaces a conventional critical value (e.g. 1.96) with a novel "OVB-adjusted" critical value that accounts for a postulated degree of omitted variable bias. These new critical values depend only on the hypothetical partial R^2 of the omitted variables with the dependent and independent variables of the OLS regression. Researchers can thus easily perform sensitivity analysis with *any standard regression software* by substituting traditional thresholds with OVB-adjusted thresholds, when testing a particular null hypothesis, or when constructing confidence intervals. We believe the extreme simplicity of implementing this approach will further aid in the widespread adoption of sensitivity analysis in applied work.

In what follows, Section 3.2 introduces the running example and provides the essential background on the main IV estimators, all of which depend upon OLS. Next, Section 3.3 refines and extends the OVB framework of Chapter 2, which not only improves the sensitivity tools for OLS, but greatly simplifies the analysis for the IV setting. Section 3.4 then develops an OVB framework for IV, first showing what can be gleaned from the first-stage and reduced-form regressions alone, then establishing the necessary OVB-type results in the Anderson-Rubin approach. Section 3.5 returns to our running example to show how these results can can be deployed in practice. Finally, we offer concluding remarks in Section 3.6. Open-source software for `R` and `Stata` implements the methods discussed in this chapter.[3]

3.2 Background

In this section we introduce the running example and use it to briefly review the required background on instrumental variables and the main approaches to IV estimation.

[3] Sensitivity analysis of the reduced form, first stage, and Anderson-Rubin regression for a specific null hypothesis can already be performed using the `R` and `Stata` package `sensemakr` [31]. Additional functionality, such as contour plots with lower and upper limits of the Anderson-Rubin confidence interval, is forthcoming.

3.2.1 Running example: estimating the returns to schooling

Ordinary least squares and the OVB problem

Many observational studies have established a positive and large association between educational achievement and earnings using regression analysis [24]. Here we consider the work of [23], which employed a sample of 3,010 individuals from the National Longitudinal Survey of Young Men (NLSYM). Considering the following multivariate linear regression

$$\text{Earnings} = \hat{\tau}_{\text{OLS,res}}\text{Education} + \boldsymbol{X}\hat{\beta}_{\text{OLS,res}} + \hat{\varepsilon}_{\text{OLS,res}} \tag{3.1}$$

where *Earnings* measures the log transformed hourly wages of the individual,[4] *Education* is an integer-valued variable indicating the completed years of education of the individual and the matrix $\boldsymbol{X}$ comprises race, experience, and a set of regional factors, Card concluded that each additional year of schooling was associated with approximately 7.5% higher wages (i.e, $\hat{\tau}_{\text{OLS,res}} \approx 0.075$) (see column "OLS" of Table 3.1).

Educational achievement, however, is not randomly assigned; perhaps individuals who obtain more education have higher wages due to other reasons, such as coming from wealthier families, or having higher levels of some unobserved characteristic, such as "ability" or "motivation." If data on these variables were available, then multivariate regression, further adjusting for such variables, would be able to capture the causal effect of educational attainment on schooling, as in

$$\text{Earnings} = \hat{\tau}_{\text{OLS}}\text{Education} + \boldsymbol{X}\hat{\beta}_{\text{OLS}} + \boldsymbol{U}\hat{\gamma}_{\text{OLS}} + \hat{\varepsilon}_{\text{OLS}} \tag{3.2}$$

where $\boldsymbol{U}$ denotes a set of variables that, along with $\boldsymbol{X}$, is sufficient to eliminate confounding concerns. Such detailed information on individuals, however, is not available, and researchers will not even agree upon which variables $\boldsymbol{U}$ are needed. In the absence of such variables,

[4] In this case, regression coefficients can be conveniently interpreted, approximately, as percent changes in earnings.

regression estimates that adjust for only a partial list of characteristics (such as $\boldsymbol{X}$) may suffer from "omitted variable bias" [6, 35] and are likely to overestimate the "true" returns to schooling.

	Dependent variable:			
	Education	Earnings (log)		
	FS	RF	OLS	IV
	(1)	(2)	(3)	(4)
Proximity	0.320^{***}	0.042^{**}		
	(0.088)	(0.018)		
Education			0.075^{***}	0.132^{**}
			(0.003)	(0.055)
Black	-0.936^{***}	-0.270^{***}	-0.199^{***}	-0.147^{***}
	(0.094)	(0.019)	(0.018)	(0.054)
SMSA	0.402^{***}	0.165^{***}	0.136^{***}	0.112^{***}
	(0.105)	(0.022)	(0.020)	(0.032)
Other covariates	yes	yes	yes	yes
Observations	3,010	3,010	3,010	3,010
R^2	0.477	0.195	0.300	0.238
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 3.1: Results of [23]. Columns show estimates and standard errors (in parenthesis) of the First Stage (FS), Reduced Form (RF), Ordinary Least Squares (OLS) and Two-Stage Least Squares (IV).

Instrumental variables as a solution to the OVB problem

Instrumental variable methods offer an alternative route to estimate the causal effect of schooling on earnings without having data on the unobserved variables $\boldsymbol{U}$. The key for such methods to work is to find a new variable (the "instrument") that changes the incentives to educational achievement, but is associated with earnings *only through* its effect on education.

To that end, [23] proposed exploiting the role of geographic differences in college accessibility. In particular, consider the variable *Proximity*, encoding an indicator of whether the individual grew up in an area with a nearby accredited 4-year college. Students who grow up far from the nearest college may face higher educational costs, discouraging them from pursuing higher level studies. Next, and most importantly, [23] argues that, conditional on the set of observed variables $\boldsymbol{X}$ (available on the NLSYM), whether one lives near a college is not itself confounded with earnings, nor does proximity to college affect earnings apart from its effect on years of education.

If we believe such assumptions hold it is possible to recover a valid estimate of the (local) average treatment effect of *Education* on *Earnings* by simply taking the ratio of two OLS coefficients, one measuring the effect of *Proximity* on *Earnings*, and another measuring the effect of *Proximity* on *Education*.[5] More precisely, consider the two OLS models

$$\text{Education} = \hat{\theta}_{\text{res}}\text{Proximity} + \boldsymbol{X}\hat{\psi}_{\text{res}} + \hat{\varepsilon}_{d,\text{res}} \tag{3.3}$$

$$\text{Earnings} = \hat{\lambda}_{\text{res}}\text{Proximity} + \boldsymbol{X}\hat{\beta}_{\text{res}} + \hat{\varepsilon}_{y,\text{res}} \tag{3.4}$$

Throughout the chapter we refer to these equations as the "first stage" (Equation 3.3) and the "reduced form" (Equation 3.4), as these are now common usage [6, 7, 83, 3].[6] The results of both regressions are also shown in Table 3.1 (columns "FS" and "RF").

The coefficient for *Proximity* on the first-stage regression, $\hat{\theta}_{\text{res}} \approx 0.32$, reveals that those who grew up near a college indeed have higher educational attainment, having completed an additional 0.32 years of education, on average. Likewise, the coefficient for *Proximity* on the reduced-form regression, $\hat{\lambda}_{\text{res}} \approx 0.042$, suggests that those who grew up near a college have

[5] This identification result requires further functional restrictions on the data-generating process, such as linearity or monotonicity. Conditions that allow a causal interpretation of the IV estimand are extensively discussed elsewhere, and will not be reviewed here. See [4], [6] and [82] for further discussion.

[6] Though now well established, these labels abuse the original meaning of the terminology, since both regressions are in their "reduced form." Equation 3.3 is called the "first stage" due to its operational role on two-stage least squares estimation, as we see next. See also [82] and [3].

4.2% higher earnings. The IV estimate is then given by the ratio of these two coefficients,

$$\hat{\tau}_{\text{res}} := \frac{\hat{\lambda}_{\text{res}}}{\hat{\theta}_{\text{res}}} \approx \frac{0.042}{0.319} \approx 0.132 \tag{3.5}$$

The value of $\hat{\tau}_{\text{res}} \approx 0.132$ suggests that, contrary to the OLS estimate of 7.5%, and perhaps surprisingly, each additional year of schooling instead raises wages by much more—13.2% (Table 3.1, column "IV").

The IV estimate itself may suffer from OVB

The previous IV estimate relies on the assumption that, conditional on $\boldsymbol{X}$, *Proximity* and *Earnings* are unconfounded, and the effect of *Proximity* on *Earnings* must go entirely through *Education*. As it is often the case, neither assumption is easy to defend in this setting. First, some of the same factors that might confound the relationship between *Education* and *Earnings* could similarly confound the relationship of *Proximity* and *Earnings* (e.g. family wealth or family connections). Second, as argued in [23], the presence of a college nearby may be associated with high school quality, which in its turn also affects earnings. Finally, other geographic confounders can make some localities likely to both have colleges nearby and lead to higher earnings. These are only coarsely conditioned on by the observed regional indicators, and residual biases may still remain.

In sum, instead of adjusting only for $\boldsymbol{X}$ as in the previous Equations 3.4 and 3.3, we should have adjusted for *both* the observed covariates $\boldsymbol{X}$ *and unobserved* covariates $\boldsymbol{W}$ as in

$$\text{Education} = \hat{\theta}\text{Proximity} + \boldsymbol{X}\hat{\psi} + \boldsymbol{W}\hat{\delta} + \hat{\varepsilon}_d \tag{3.6}$$

$$\text{Earnings} = \hat{\lambda}\text{Proximity} + \boldsymbol{X}\hat{\beta} + \boldsymbol{W}\hat{\gamma} + \hat{\varepsilon}_y \tag{3.7}$$

Where $\boldsymbol{W}$ stands for all unobserved factors necessary to make *Proximity* a valid instrument for the effect of *Education* on *Earnings* (e.g, *Family Wealth*, *High School Quality*, *Place of*

Residence, etc). The IV estimate we wished we had is then given by

$$\hat{\tau} := \frac{\hat{\lambda}}{\hat{\theta}} \tag{3.8}$$

Our previous estimate $\hat{\tau}_{\text{res}}$ deviates from the target estimate $\hat{\tau}$, but how badly? How strong would the omitted variables $\boldsymbol{W}$ have to be so that it would change our research conclusions? To develop a precise algebraic answer to this question, we must first review the mechanics of the main approaches to IV estimation.

3.2.2 The mechanics of IV estimation

Let the random variable Y_i denote the outcome, D_i the treatment, Z_i the instrumental variable, $\boldsymbol{X}_i = [\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{ip}]$ a vector of p observed covariates, and $\boldsymbol{W}_i = [\boldsymbol{W}_{i1}, \ldots, \boldsymbol{W}_{il}]$ a vector of l unobserved covariates for an individual. The target quantity of IV estimation consists of a ratio of two *population* regression coefficients,

$$\tau := \frac{\lambda}{\theta} \tag{3.9}$$

where θ is the population regression coefficient of Z_i on D_i (the first stage) and λ the population regression coefficient of Z_i on Y_i (the reduced form), both adjusting for $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$. We call the ratio τ the *IV estimand*. Here we briefly review the commonly used approaches to make inferences regarding this ratio.

3.2.2.1 Indirect Least Squares and Two-Stage Least Squares

Throughout the chapter we consider exact algebraic results that holds for sample estimates. Denote by Y the $(n \times 1)$ *vector* of the outcome of interest with n observations; by D the $(n \times 1)$ treatment vector; by Z the $(n \times 1)$ vector of the instrument; by $\boldsymbol{X}$ an $(n \times p)$ *matrix* of observed covariates (including a constant), and by $\boldsymbol{W}$ an $(n \times l)$ matrix of *unobserved* covariates.

Indirect Least Squares. The first and perhaps most straightforward approach to instrumental variable estimation was outlined above: run two OLS models capturing the effect of the instrument on the treatment (first stage) and the effect of the instrument on the outcome (reduced form),

$$\textbf{First stage:} \quad D = \hat{\theta} Z + \boldsymbol{X}\hat{\psi} + \boldsymbol{W}\hat{\delta} + \hat{\varepsilon}_d \tag{3.10}$$

$$\textbf{Reduced form:} \quad Y = \hat{\lambda} Z + \boldsymbol{X}\hat{\beta} + \boldsymbol{W}\hat{\gamma} + \hat{\varepsilon}_y \tag{3.11}$$

Where $\hat{\theta}$, $\hat{\psi}$ and $\hat{\delta}$ are the OLS estimates of the regression of D on Z, $\boldsymbol{X}$ and $\boldsymbol{W}$, and $\hat{\varepsilon}_d$ its corresponding residuals; analogously, $\hat{\lambda}$, $\hat{\beta}$ and $\hat{\gamma}$ are the OLS estimates of the regression of Y on Z, $\boldsymbol{X}$ and $\boldsymbol{W}$, and $\hat{\varepsilon}_y$ its corresponding residuals. The estimator for τ is constructed by simply using the plug-in principle and taking the ratio of $\hat{\lambda}$ and $\hat{\theta}$

$$\hat{\tau}_{\text{ILS}} := \frac{\hat{\lambda}}{\hat{\theta}} \tag{3.12}$$

The ratio $\hat{\tau}_{\text{ILS}}$ may be called the *indirect least squares* (ILS) estimator, or the "ratio of coefficients" estimator. Inference in the ILS framework can be performed using the delta-method, resulting in the *estimated* variance

$$\widehat{\text{var}}(\hat{\tau}_{\text{ILS}}) := \frac{1}{\hat{\theta}^2}\left(\widehat{\text{var}}(\hat{\lambda}) + \hat{\tau}_{\text{ILS}}^2\widehat{\text{var}}(\hat{\theta}) - 2\hat{\tau}_{\text{ILS}}\widehat{\text{cov}}(\hat{\lambda},\hat{\theta})\right) \tag{3.13}$$

Where $\widehat{\text{var}}(\hat{\lambda})$, $\widehat{\text{var}}(\hat{\theta})$ and $\widehat{\text{cov}}(\hat{\lambda},\hat{\theta})$ are the usual OLS variance and covariance estimates (see appendix).

Two-Stage Least Squares. A closely related approach for instrumental variable estimation is denoted by "two-stage least squares" (2SLS). As its name suggests, this involves two nested steps of OLS estimation: a first-stage regression given by Equation 3.10 to produce fitted values for the treatment ($\widehat{D}$), then regressing the outcome on these fitted values,

$$\textbf{Second stage:} \quad Y = \hat{\tau}_{\text{2SLS}}\widehat{D} + \boldsymbol{X}\hat{\beta}_{\text{2SLS}} + \boldsymbol{W}\hat{\gamma}_{\text{2SLS}} + \hat{\varepsilon}_{\text{2SLS}} \tag{3.14}$$

The 2SLS estimate corresponds to the coefficient $\hat{\tau}_{\text{2SLS}}$ in Equation 3.14, called the "second-stage" regression. By the Frisch-Waugh-Lovell (FWL) theorem (63, 93, 94), one can readily show that $\hat{\tau}_{\text{2SLS}}$ and $\hat{\tau}_{\text{ILS}}$ are numerically identical,

$$\hat{\tau}_{\text{2SLS}} = \frac{\text{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}}, \widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})}{\text{var}(\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})} = \frac{\hat{\theta} \times \text{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}}, Z^{\perp \boldsymbol{X},\boldsymbol{W}})}{\hat{\theta}^2 \times \text{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} = \frac{\hat{\lambda}}{\hat{\theta}} \tag{3.15}$$

Where $Y^{\perp \boldsymbol{X},\boldsymbol{W}}$,$\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}}$ and $D^{\perp \boldsymbol{X},\boldsymbol{W}}$ denote the variables Y, $\widehat{D}$ and D after removing the components linearly explained by $\boldsymbol{X}$ and $\boldsymbol{W}$, and cov$(\cdot)$ and var$(\cdot)$ denote the *sample* covariance and variance of those variables. As with ILS, inference in 2SLS is performed by resorting to the asymptotic normality of the ratio, with estimated variance

$$\widehat{\text{var}}(\hat{\tau}_{\text{2SLS}}) := \frac{\text{var}(Y^{\perp \boldsymbol{X},\boldsymbol{W}} - \hat{\tau}_{\text{2SLS}} D^{\perp \boldsymbol{X},\boldsymbol{W}})}{\text{var}(\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})} \times \text{df}^{-1} \tag{3.16}$$

Where df denotes the appropriate degrees of freedom. Using the FWL theorem one can further show that $\widehat{\text{var}}(\hat{\tau}_{\text{2SLS}})$ and $\widehat{\text{var}}(\hat{\tau}_{\text{ILS}})$ are also numerically identical (see appendix).

3.2.2.2 Anderson-Rubin regression and Fieller's theorem

The methods of ILS and 2SLS make use of a normal approximation to the sampling distribution of the ratio $\hat{\lambda}/\hat{\theta}$, which may prove unreliable when θ is "close" to zero, relative to the sampling variability of $\hat{\theta}$—this is known as the "weak instrument" problem. Two alternatives that allow constructing confidence intervals with correct coverage, regardless of the "strength" of the first stage, are the proposals of [2] and [52] (e.g. see 3).

Anderson-Rubin. The Anderson-Rubin approach starts by creating the random variable $Y_{\tau_0} := Y - \tau_0 D$ in which we subtract from Y a "putative" causal effect of D, namely, τ_0. If Z is a valid instrument, under the null hypothesis $H_0 : \tau = \tau_0$, we should not see an association between Y_{τ_0} and Z, conditional on $\boldsymbol{X}$ and $\boldsymbol{W}$. In other words, if we run the OLS model

$$\textbf{Anderson-Rubin:} \quad Y_{\tau_0} = \hat{\phi}_{\tau_0} Z + \boldsymbol{X}\hat{\beta}_{\tau_0} + \boldsymbol{W}\hat{\gamma}_{\tau_0} + \hat{\varepsilon}_{\tau_0} \tag{3.17}$$

we should find that $\hat{\phi}_{\tau_0}$ is equal to zero, but for sampling variation. To test the null hypothesis $H_0 : \phi_{\tau_0} = 0$ in the Anderson-Rubin regression is thus equivalent to test the null hypothesis $H_0 : \tau = \tau_0$. The $1 - \alpha$ confidence interval is constructed by collecting all values τ_0 such that the null hypothesis $H_0 : \phi_{\tau_0} = 0$ is not rejected at the chosen significance level α:

$$\mathrm{CI}_{1-\alpha}(\tau) := \{\tau_0;\ t^2_{\hat{\phi}_{\tau_0}} \leq t^{*2}_{\alpha,\mathrm{df}}\} \tag{3.18}$$

Where $t_{\hat{\phi}_{\tau_0}}$ is the t-value of the coefficient $\hat{\phi}_{\tau_0}$, and $t^*_{\alpha,\mathrm{df}}$ the usual α level critical threshold for the t statistic, with the appropriate degrees of freedom. It is also convenient to define the point estimate $\hat{\tau}_{\mathrm{AR}}$ as the value τ_0 which makes $\hat{\phi}_{\tau_0}$ exactly equal to zero

$$\hat{\tau}_{\mathrm{AR}} := \{\tau_0;\ \hat{\phi}_{\tau_0} = 0\} \tag{3.19}$$

By the FWL theorem, we can write $\hat{\phi}_{\tau_0}$ as a linear combination of $\hat{\lambda}$ and $\hat{\theta}$,

$$\hat{\phi}_{\tau_0} = \frac{\mathrm{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}} - \tau_0 D^{\perp \boldsymbol{X},\boldsymbol{W}}, Z^{\perp \boldsymbol{X},\boldsymbol{W}})}{\mathrm{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} = \hat{\lambda} - \tau_0 \hat{\theta} \tag{3.20}$$

Thus resulting in $\hat{\tau}_{AR} = \frac{\hat{\lambda}}{\hat{\theta}}$, a point estimate numerically identical to the previous estimators.

Fieller's theorem. The connection between Fieller's theorem and the Anderson-Rubin approach follows from Equation 3.20. The central test statistic of Fieller's theorem is precisely the linear combination $\hat{\phi}_{\tau_0} = \hat{\lambda} - \tau_0 \hat{\theta}$. Under the null hypothesis $H_0 : \tau = \tau_0$, if the estimators $\hat{\lambda}$ and $\hat{\theta}$ are asymptotically normal, it follows that $\hat{\phi}_{\tau_0}$ is also asymptotically normal with mean zero, and estimated variance

$$\widehat{\mathrm{var}}(\hat{\phi}_{\tau_0}) := \widehat{\mathrm{var}}(\hat{\lambda}) + \tau_0^2 \widehat{\mathrm{var}}(\hat{\theta}) - 2\tau_0 \widehat{\mathrm{cov}}(\hat{\lambda}, \hat{\theta}) \tag{3.21}$$

Confidence intervals are then constructed exactly as in Equation 3.18, and the two approaches are numerically identical.

3.2.3 Problem statement

As we have seen, all main approaches for IV estimation result in the same point estimate—the ratio of the reduced-form and first-stage regression coefficients. They differ only in how to perform inference, with ILS/2SLS resorting to the asymptotic normality of the ratio estimator, and the Anderson-Rubin/Fieller approach inverting the test of the linear combination of both coefficients.

	Restricted IV regressions	Full IV regressions
First stage	$D = \hat{\theta}_{\text{res}} Z + \boldsymbol{X}\hat{\psi}_{\text{res}} + \hat{\varepsilon}_{d,\text{res}}$	$D = \hat{\theta} Z + \boldsymbol{X}\hat{\psi} + \boldsymbol{W}\hat{\delta} + \hat{\varepsilon}_{d}$
Reduced form	$Y = \hat{\lambda}_{\text{res}} Z + \boldsymbol{X}\hat{\beta}_{\text{res}} + \hat{\varepsilon}_{y,\text{res}}$	$Y = \hat{\lambda} Z + \boldsymbol{X}\hat{\beta} + \boldsymbol{W}\hat{\gamma} + \hat{\varepsilon}_{y}$
Anderson-Rubin	$Y_{\tau_0} = \hat{\phi}_{\tau_0,\text{res}} Z + \boldsymbol{X}\hat{\beta}_{\tau_0,\text{res}} + \hat{\varepsilon}_{\tau_0,\text{res}}$	$Y_{\tau_0} = \hat{\phi}_{\tau_0} Z + \boldsymbol{X}\hat{\beta}_{\tau_0} + \boldsymbol{W}\hat{\gamma}_{\tau_0} + \hat{\varepsilon}_{\tau_0}$

Table 3.2: The omitted variable bias problem for instrumental variable regressions.

The regression equations discussed in Section 3.2.2, summarized in the third column of Table 3.2, stand for the IV regressions our analyst *wished* she had run, adjusting for both $\boldsymbol{X}$ and $\boldsymbol{W}$. However, since $\boldsymbol{W}$ is *unobserved*, the investigator is forced to run instead the restricted models in the second column of Table 3.2. Our task is thus to characterize how point estimates and confidence intervals for the IV estimate, given by these regressions, would have changed due to the inclusion of $\boldsymbol{W}$. Since, at their core, all these IV approaches rely on OLS estimation, we should be able to leverage all OVB tools for OLS for examining the sensitivity of IV.

A note on identification with instrumental variables

Before proceeding, it is worth making a brief note on the identification of causal effects using instrumental variables. There are many different sets of assumptions that allow different causal interpretations of the IV estimand given by Equation 3.9 [4, 20, 109, 135]. The causal diagram of Figure 3.1 shows some of the most used "canonical" models illustrating the main traditional assumptions of IV. Equivalent assumptions can be articulated in the potential outcomes framework [109, 135]. Beyond those assumptions of exclusion and independence restrictions,

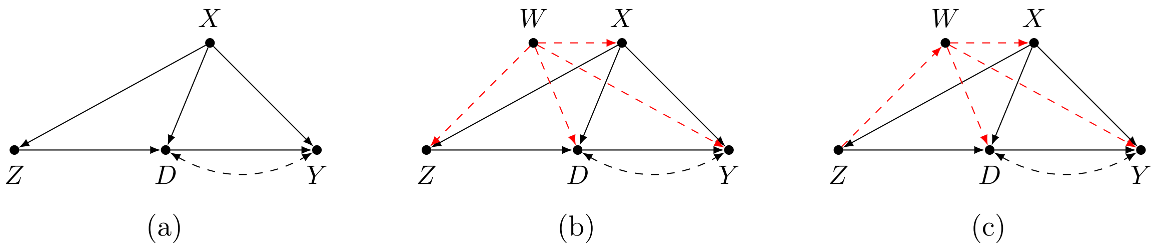


Figure 3.1: Causal diagrams illustrating the traditional IV assumptions. In Figure 3.1a, X is sufficient for rendering Z a valid instrumental variable. In Figures 3.1b and 3.1c, however, W is also needed to render Z a valid IV (in Figure 3.1b W is a confounder of the instrument-outcome relationship, whereas in Figure 3.1c W is a side-effect of the instrument). Graphically, conditional on a set of covariates $\{X, W\}$, a variable Z is a valid instrument for the causal effect of a treatment D on an outcome Y, if the set $\{X, W\}$ blocks all paths from Z to Y on the graph where the edge $D \rightarrow Y$ is removed [20].

some functional constraint is also needed for point-identification. For instance, under certain assumptions of effect homogeneity (e.g., linearity), the IV estimand can be interpreted as the average treatment effect; another widely used example is the binary setting with the assumption of monotonicity, in which case the IV estimand can be interpreted as a local average treatment effect [4, 6, 135]. Here we do not commit to a specific causal interpretation, and simply assume the researcher is interested in the IV estimand of Equation 3.9, adjusting for both $\boldsymbol{X}$ and $\boldsymbol{W}$. All sensitivity results we present here are thus valid for *any* set of IV assumptions, so long as the resulting estimand is still given by Equation 3.9.

3.3 Omitted variable bias with the partial R^2 parameterization

In this section, we extend the results of Chapter 2 regarding the partial R^2 parameterization of the OVB formula for OLS. In particular, we introduce the notion of *OVB-adjusted* critical values, and show how sensitivity analysis can be performed by simply substituting traditional critical values with the adjusted ones. We also introduce the idea of a set of compatible inferences given bounds on the strength of confounding, and formalize sensitivity statistics for routine reporting as answering an inverse question regarding those sets. These extensions are not only useful for the sensitivity of OLS estimates themselves, but will greatly simplify the generalization of these results to the IV setting in Section 3.4. To fix ideas, here we discuss the OVB framework in the context of the reduced-form regression coefficient, but the reader should have in mind that all results presented here are algebraic, and hold for *any* OLS estimate.

3.3.1 Sensitivity in an omitted variable bias framework

The OVB framework starts with a target coefficient obtained from a *full* regression equation that the analyst wished she could have estimated (such as those in the third column of Table 3.2). For concreteness, suppose we are interested in the coefficient $\hat{\lambda}$ of the regression equation of the outcome Y on the instrument Z, adjusting for a set of observed covariates $\boldsymbol{X}$ and a single *unobserved* covariate W (we generalize to multivariate W below),

$$Y = \hat{\lambda} Z + \boldsymbol{X}\hat{\beta} + \hat{\gamma} W + \hat{\varepsilon}_y \tag{3.22}$$

However, when W is unobserved, estimating the full regression equation is infeasible. Instead, the investigator is forced to estimate the *restricted* model given by

$$Y = \hat{\lambda}_{\text{res}} Z + \boldsymbol{X}\hat{\beta}_{\text{res}} + \hat{\varepsilon}_{y,\text{res}} \tag{3.23}$$

Where $\hat{\lambda}_{\text{res}}$ and $\hat{\beta}_{\text{res}}$ are the coefficients of the restricted OLS adjusting for Z and $\boldsymbol{X}$ alone, and $\hat{\varepsilon}_{y,\text{res}}$ its corresponding residual. The OVB framework seeks to answer the following question: how do the inferences for λ_{res} from the restricted OLS model (omitting W), compare with the inferences for λ from the full OLS model (adjusting for W)?

3.3.1.1 Adjusted estimates and standard errors

Let $R^2_{Y\sim W|Z,\boldsymbol{X}}$ denote the partial R^2 of W with Y, after controlling for Z and $\boldsymbol{X}$, and let $R^2_{Z\sim W|\boldsymbol{X}}$ denote the partial R^2 of W with Z after adjusting for $\boldsymbol{X}$. Given the estimates of the restricted model, $\hat{\lambda}_{\text{res}}$ and $\widehat{\text{se}}(\hat{\lambda}_{\text{res}})$, the values $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$ are sufficient to recover $\hat{\lambda}$ and $\widehat{\text{se}}(\hat{\lambda})$ as we have seen in Chapter 2. More precisely, define $\widehat{\text{bias}}(\lambda) := \hat{\lambda}_{\text{res}} - \hat{\lambda}$

as the difference between the restricted estimate and the full estimate. We then have,

$$|\widehat{\text{bias}}(\lambda)| = \sqrt{\frac{R^2_{Y\sim W|Z,\boldsymbol{X}} R^2_{Z\sim W|\boldsymbol{X}}}{1 - R^2_{Z\sim W|\boldsymbol{X}}} \text{df}} \times \widehat{\text{se}}(\hat{\lambda}_{\text{res}}) = \text{BF}\sqrt{\text{df}} \times \widehat{\text{se}}(\hat{\lambda}_{\text{res}}) \tag{3.24}$$

Where hereafter df $= n - p - 1$ stands for the degrees of freedom of the restricted model *actually run*. For notational convenience, and to aid interpretation, we define the term

$$\text{BF} := \sqrt{\frac{R^2_{Y\sim W|Z,\boldsymbol{X}} R^2_{Z\sim W|\boldsymbol{X}}}{1 - R^2_{Z\sim W|\boldsymbol{X}}}} \tag{3.25}$$

as the "bias factor" of W, which is the part of the bias solely determined by $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$. Likewise, the standard error of the full model can be recovered with

$$\widehat{\text{se}}(\hat{\lambda}) = \sqrt{\frac{1 - R^2_{Y\sim W|Z,\boldsymbol{X}}}{1 - R^2_{Z\sim W|\boldsymbol{X}}} \left(\frac{\text{df}}{\text{df} - 1}\right)} \times \widehat{\text{se}}(\hat{\lambda}_{\text{res}}) = \text{SEF}\sqrt{\text{df}/(\text{df} - 1)} \times \widehat{\text{se}}(\hat{\lambda}_{\text{res}}) \tag{3.26}$$

Where again, for convenience, we define

$$\text{SEF} := \sqrt{\frac{1 - R^2_{Y\sim W|Z,\boldsymbol{X}}}{1 - R^2_{Z\sim W|\boldsymbol{X}}}} \tag{3.27}$$

as the "standard error factor" of W, summarizing the factor of the adjusted standard error which is solely determined by the sensitivity parameters $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$. Note again that SEF consists of the square-root of the product of the familiar "variance inflation factor," $1/\left(1 - R^2_{Z\sim W|\boldsymbol{X}}\right)$ and what could be labeled the "variance reduction factor," $1 - R^2_{Y\sim W|Z,\boldsymbol{X}}$, as discussed in Section 2.4.2. As we have seen, although simple, Equations 3.24 and 3.26 form the basis of a rich set of sensitivity exercises regarding point estimates, standard errors and t-values in terms of sensitivity parameters $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$.

Multiple unobserved variables. For simplicity of exposition, throughout the chapter we usually refer to a single omitted variable W. These results, however, can be used for performing sensitivity analyses considering multiple omitted variables $\boldsymbol{W} = [W_1, W_2, \ldots, W_n]$, and thus also non-linearities and functional form misspecification of observed variables. In such cases, barring an adjustment in the degrees of freedom, the equations are conservative, and reveal the maximum bias a multivariate $\boldsymbol{W}$ with such pair of partial R^2 values could cause, as discussed in Section 2.4.5.

3.3.1.2 Adjusted lower and upper limits of confidence intervals

We now closely examine how the confidence interval of a regression coefficient changes due to the inclusion of W. Traditional confidence intervals account for sampling uncertainty, and are constructed by multiplying the standard error of the coefficient by a critical value (for example, in large samples, 1.96 for a 95% confidence level). We show that replacing this traditional critical value with an *OVB-adjusted critical value*, which we introduce here, accounts for both sampling uncertainty and systematic biases due to the omission of W. Although simple, this perspective will prove useful for deriving and understanding OVB-type results for OLS in general, and for instrumental variables in particular, such as in the Anderson-Rubin approach of Section 3.4.

Specifically, let $t^*_{\alpha,\mathrm{df}-1}$ denote the critical value for a standard t-test with significance level α and $\mathrm{df}-1$ degrees of freedom. Now let $\mathrm{LL}_{1-\alpha}(\lambda)$ be the lower limit and $\mathrm{UL}_{1-\alpha}(\lambda)$ be the upper limit of a $1-\alpha$ confidence interval for λ in the full model, i.e.,

$$\mathrm{LL}_{1-\alpha}(\lambda) := \hat{\lambda} - t^*_{\alpha,\mathrm{df}-1} \times \widehat{\mathrm{se}}(\hat{\lambda}), \quad \mathrm{UL}_{1-\alpha}(\lambda) := \hat{\lambda} + t^*_{\alpha,\mathrm{df}-1} \times \widehat{\mathrm{se}}(\hat{\lambda}), \tag{3.28}$$

Considering the direction of the bias that further reduces the lower limit, or, alternatively, a direction that further increases the upper limit, Equations 3.24 and 3.26 imply that both quantities can be written as a function of the restricted estimates and a new multiplier (see appendix)

$$\mathrm{LL}_{1-\alpha}(\lambda) = \hat{\lambda}_{\mathrm{res}} - t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}), \quad \mathrm{UL}_{1-\alpha}(\lambda) = \hat{\lambda}_{\mathrm{res}} + t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) \tag{3.29}$$

where $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ stands for the *OVB-adjusted critical value*

$$t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} := \mathrm{SEF}\sqrt{\mathrm{df}/(\mathrm{df}-1)} \times t^{*}_{\alpha,\mathrm{df}-1} + \mathrm{BF}\sqrt{\mathrm{df}}. \tag{3.30}$$

The subscript $\boldsymbol{R}^2 = \{R^2_{Y\sim W|Z,\boldsymbol{X}}, R^2_{Z\sim W|\boldsymbol{X}}\}$ conveys the fact that $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ depends on both sensitivity parameters. The adjusted critical value $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ *uniquely determines* the extreme points of the confidence interval for λ that one could obtain after adjusting for an omitted variable W with a given pair of partial R^2. Equivalently, given any hypothetical strength of W, to test the general null hypothesis of a change of $(100 \times q^*)\%$ of the current estimate $\hat{\lambda}_{\mathrm{res}}$ at the α level, it suffices to rescale the original t-value by q^* and compare this to the adjusted critical threshold $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$.[7]

3.3.1.3 Compatible inferences given bounds on partial R^2

Given hypothetical values for $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$, the previous results allow us to determine the exact changes in inference regarding a parameter of interest due to the inclusion of W with such strength. Often, however, the analyst does not know the exact strength of omitted variables, and wishes to investigate the *worst* possible inferences that could be induced by a W with bounded strength, for instance, $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq R^{2\,\max}_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\,\max}_{Z\sim W|\boldsymbol{X}}$. That is, we wish to find the maximum adjusted critical value due to an omitted variable W with *at most* such strength. Writing $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ as a function of the sensitivity parameters $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$, we solve the maximization problem

$$\max_{R^2_{Y\sim W|Z,\boldsymbol{X}},R^2_{Z\sim W|\boldsymbol{X}}} t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \quad \text{s.t.} \quad R^2_{Y\sim W|Z,\boldsymbol{X}} \leq R^{2\,\max}_{Y\sim W|Z,\boldsymbol{X}}, \quad R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\,\max}_{Z\sim W|\boldsymbol{X}} \tag{3.31}$$

[7] For a numerical example of an adjusted critical value, consider a case with 100 degrees of freedom and a significance level of $\alpha = 5\%$. The traditional critical value, assuming no omitted variables, is $t^{*}_{.05,100} \approx 1.98$. If we now allow for an omitted variable with strength given by $R^2_{Y\sim W|Z,\boldsymbol{X}} = R^2_{Z\sim W|\boldsymbol{X}} = .1$, this leads to an increased OVB-adjusted critical value of $t^{\dagger}_{.05,100,.1,.1} \approx 3.05$. Further note $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ *increases* the larger the sample size—for instance, if the degrees of freedom were instead 1,000, the adjusted critical value would increase to approximately 5.30.

Note that, although this maximum is often reached at the extrema of both coordinates, this is not always the case. Due to the variance reduction factor, increasing $R^2_{Y\sim W|Z,\boldsymbol{X}}$ may reduce the standard error more than enough to compensate for the increase in bias, resulting in tighter confidence intervals. Denoting the solution to the optimization problem in expression (3.31) as $t^{\dagger\,\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$, the *most extreme possible* lower and upper limits after adjusting for W are given by

$$\mathrm{LL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda) = \hat{\lambda}_{\mathrm{res}} - t^{\dagger\,\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}), \quad \mathrm{UL}^{\max}_{1-\alpha,\boldsymbol{R}^2} = \hat{\lambda}_{\mathrm{res}} + t^{\dagger\,\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) \tag{3.32}$$

The interval composed of such limits,

$$\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda) = \left[\mathrm{LL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda), \quad \mathrm{UL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)\right] \tag{3.33}$$

retrieves all inferences for λ which are compatible with an omitted variable with such strengths. In other words, without imposing further constraints on W, for any value λ_0 inside $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$, we can find a W such that $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq R^{2\,\max}_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\,\max}_{Z\sim W|\boldsymbol{X}}$ and the confidence interval for λ after adjusting for W includes λ_0. Moreover, if the true partial R^2 of W lies within the posited bounds, then $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$ is the union of all confidence intervals that would be obtained by including an omitted variable with that strength or less, and thus constitutes itself a confidence interval with *at least* $1-\alpha$ coverage (provided, of course, our "target" confidence interval adjusting for W has nominal coverage).

3.3.2 Sensitivity statistics for routine reporting

Widespread adoption of sensitivity analysis benefits from simple and interpretable statistics that quickly convey the overall robustness of an estimate. To that end, in Chapter 2 we proposed two sensitivity statistics for routine reporting: (i) the partial R^2 of Z with Y, $R^2_{Y\sim Z|\boldsymbol{X}}$; and, (ii) the *robustness value* (RV). Here we generalize the notion of a partial R^2 as a measure of robustness to extreme scenarios, by introducing the *extreme robustness value* (XRV), for which the partial R^2 is a special case. We also recast these sensitivity statistics as a solution to an "inverse" question regarding the interval of compatible inferences, $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$—that is, given a threshold of inference for λ deemed to be of scientific importance (say, zero), what is the *minimum* strength of the sensitivity parameters $\boldsymbol{R}^2$ that could lead $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$ to include such threshold? This framework facilitates extending these metrics to other contexts, in particular to the IV setting, as we show in Section 3.4.2.3.

3.3.2.1 The extreme robustness value

One benefit of the partial R^2 parameterization is that the parameter $R^2_{Y\sim W|Z,\boldsymbol{X}}$ can be left completely unconstrained; i.e, in the optimization problem of expression 3.31, one can set the bound for $R^2_{Y\sim W|Z,\boldsymbol{X}}$ to its trivial bound of 1, and this still results in non-trivial bounds on the set of possible inferences. This leads to our first inverse question: what is the *bare minimum* strength of association of the omitted variable W with Z that could bring its estimated coefficient to a region where it is no longer statistically different than zero (or another threshold of interest)?

To answer this question, we can see $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$ as a function of the bound $R^{2\,\max}_{Z\sim W|\boldsymbol{X}}$ alone, obtained from maximizing the adjusted critical value in expression 3.31 where: (i) the parameter $R^2_{Y\sim W|Z,\boldsymbol{X}}$ is left completely unconstrained (i.e, $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq 1$); and, (ii) the parameter $R^2_{Z\sim W|\boldsymbol{X}}$ is bounded by XRV (i.e, $R^{2\,\max}_{Z\sim W|\boldsymbol{X}} \leq \mathrm{XRV}$). The *Extreme Robustness Value* $\mathrm{XRV}_{q^*,\alpha}(\lambda)$ is defined as the greatest lower bound XRV such that the null hypothesis that a change of $(100 \times q^*)\%$ of the original estimate, $H_0 : \lambda = (1-q^*)\hat{\lambda}_{\text{res}}$, is not rejected at the α level,

$$\mathrm{XRV}_{q^*,\alpha}(\lambda) := \inf\left\{\mathrm{XRV};\ (1-q^*)\hat{\lambda}_{\text{res}} \in \mathrm{CI}^{\max}_{1-\alpha,1,\mathrm{XRV}}(\lambda)\right\} \tag{3.34}$$

The solution to this problem gives,

$$\mathrm{XRV}_{q^*,\alpha}(\lambda) = \begin{cases} 0, & \text{if } \; f_{q^*}(\lambda) \leq f^*_{\alpha,\mathrm{df}-1} \\ \dfrac{f^2_{q^*}(\lambda) - f^{*2}_{\alpha,\mathrm{df}-1}}{1 + f^2_{q^*}(\lambda)}, & \text{otherwise.} \end{cases} \tag{3.35}$$

Where $f_{q^*}(\lambda) := q^*|f_{Y\sim Z|\boldsymbol{X}}|$ (here $f_{Y\sim Z|\boldsymbol{X}}$ stands for the partial Cohen's f and we define the critical threshold $f^*_{\alpha,\mathrm{df}\,-1} := t^*_{\alpha,\mathrm{df}\,-1}/\sqrt{\mathrm{df}\,-1}$).[8] Note $\mathrm{XRV}_{q^*,\alpha}(\lambda)$ can be interpreted as an "adjusted partial R^2" of Z with Y. To see why, let us first consider the case of the minimal strength to bring the point estimate ($\alpha = 1$) to exactly zero ($q^* = 1$). We then have that $f^*_{\alpha=1,\mathrm{df}\,-1} = 0$ and $f^2_{q^*=1}(\lambda) = f^2_{Y\sim Z|\boldsymbol{X}}$, resulting in

$$\mathrm{XRV}_{q^*=1,\alpha=1}(\lambda) = \frac{f^2_{Y\sim Z|\boldsymbol{X}}}{1 + f^2_{Y\sim Z|\boldsymbol{X}}} = R^2_{Y\sim Z|\boldsymbol{X}} \tag{3.36}$$

This recovers the result of Section 2.4.3, and shows that, for an omitted variable W to bring down the estimated coefficient to zero, it needs to explain at least as much residual variation of Z, as Z explains of Y. For the general case, we simply perform two adjustments that dampens the "raw" partial R^2 of Z with Y. First we adjust it by the proportion of reduction deemed to be problematic q^* through $f_{q^*} = q^*|f_{Y\sim Z|\boldsymbol{X}}|$; next, we subtract the threshold for which statistical significance is lost at the α level (via $f^{*2}_{\alpha,\mathrm{df}-1}$).

The extreme robustness value establishes thus the equivalent of a "Cornfield condition" [42] for OLS estimates, and delineates the bare minimum strength of omitted variables necessary to overturn a certain conclusion—if W cannot explain at least $\mathrm{XRV}_{q^*,\alpha}(\lambda)$ of the residual variation of Z, then such variable *is not* strong enough to bring about a change of $(100 \times q^*)\%$ on the original estimate, at the significance level of α, regardless of its association with Y.

[8] Cohen's f^2 can be written as $f^2 = R^2/(1 - R^2)$.

3.3.2.2 The robustness value

Placing no constraints on the association of the omitted variable W with Y may be too conservative an exercise. An alternative measure of robustness of the OLS estimate is to consider the minimal strength of association that the omitted variable needs to have, *both* with Z and Y, so that a $1-\alpha$ confidence interval for λ will include a change of $(100 \times q^*)\%$ of the current restricted estimate.

Write $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$ as a function of both bounds varying simultaneously, that is, construct $\mathrm{CI}^{\max}_{1-\alpha,\mathrm{RV},\mathrm{RV}}(\lambda)$ by maximizing the adjusted critical value with bounds given by $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq \mathrm{RV}$ and $R^2_{Z\sim W|\boldsymbol{X}} \leq \mathrm{RV}$. The *Robustness Value* $\mathrm{RV}_{q^*,\alpha}(\lambda)$ for not rejecting the null hypothesis that $H_0 : \lambda = (1-q^*)\hat{\lambda}_{\mathrm{res}}$, at the significance level α, is defined as

$$\mathrm{RV}_{q^*,\alpha}(\lambda) := \inf\left\{\mathrm{RV};\ (1-q^*)\hat{\lambda}_{\mathrm{res}} \in \mathrm{CI}^{\max}_{1-\alpha,\mathrm{RV},\mathrm{RV}}(\lambda)\right\} \tag{3.37}$$

We then have that,

$$\mathrm{RV}_{q^*,\alpha}(\lambda) = \begin{cases} 0, & \text{if } \ f_{q^*}(\lambda) \leq f^*_{\alpha,\mathrm{df}-1} \\ \frac{1}{2}\left(\sqrt{f^4_{q^*,\alpha}(\lambda) + 4f^2_{q^*,\alpha}(\lambda)} - f^2_{q^*,\alpha}(\lambda)\right), & \text{if } \ f^*_{\alpha,\mathrm{df}-1} < f_{q^*}(\lambda) < f^{*-1}_{\alpha,\mathrm{df}-1} \\ \mathrm{XRV}_{q^*,\alpha}(\lambda), & \text{otherwise.} \end{cases} \tag{3.38}$$

Where $f_{q^*,\alpha}(\lambda) := q^*|f_{Y\sim Z|\boldsymbol{X}}| - f^*_{\alpha,\mathrm{df}\,-1}$. In the appendix we show the conditions of Equation 3.38 are equivalent to those we had previously derived, with the advantage of being simpler to verify. The first case occurs when the confidence interval already includes $(1-q^*)\hat{\lambda}_{\mathrm{res}}$ or the mere change of one degree of freedom achieves this. The second case occurs when both associations of W reach the bound. Finally, in the last case the solution is an interior point—this happens when the bound is large enough such that the constraint on the association with the outcome is not binding; in this case the RV reduces to the XRV.

The robustness value offers a simple interpretable measure that summarizes the strength of omitted variables necessary to change the estimate in problematic ways. If W explains $\mathrm{RV}_{q^*,\alpha}(\lambda)$ of the residual variance of both Z and Y, then such variable is sufficiently strong to bring about a $(100 \times q)\%$ change in the estimate at the significance level of α, while any omitted variable that does not explain $\mathrm{RV}_{q^*,\alpha}(\lambda)$ of the residual variance, neither of Z nor of Y, is not sufficiently strong to do so.

A visual depiction of the RV and XRV

Visually depicting the RV and the XRV in a sensitivity contour plot may be helpful. Consider Figure 3.2. The horizontal axis describes $R^2_{Z\sim W|\boldsymbol{X}}$ and the vertical axis describes $R^2_{Y\sim W|Z,\boldsymbol{X}}$. The contour lines show the adjusted t-value for testing the null hypothesis of zero effect for the reduced form regression (of Table 3.1), had we adjusted for W with such hypothetical strength (considering that adjustment reduces the t-value). The red dashed line shows a critical contour of interest, such as statistical significance at the $\alpha = 0.05$ level. The RV (when both values reach their bounds) summarizes the point of equal values on both axis of the critical contour, whereas the XRV summarizes the vertical line tangent to the critical contour, which will never be crossed.

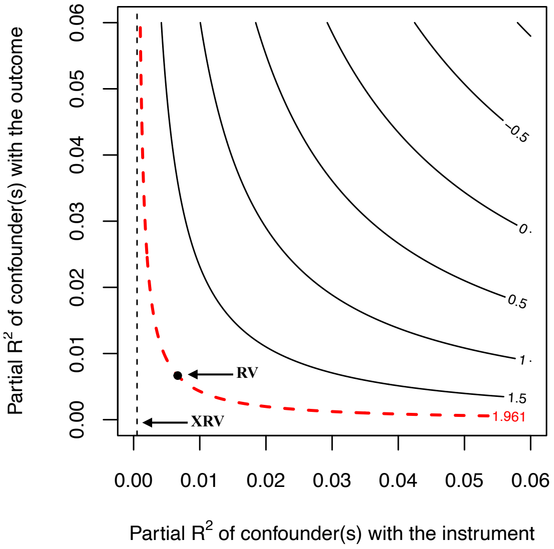


Figure 3.2: Sensitivity contours of the reduced form of [23] depicting the RV and the XRV.

3.3.3 Bounding the strength of the omitted variable using observed covariates

One further result is required before turning to the sensitivity of IV estimates. Let X_j be a specific covariate of the set $\boldsymbol{X}$, and define

$$k_Z := \frac{R^2_{Z\sim W|\boldsymbol{X}_{-j}}}{R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}}, \qquad k_Y := \frac{R^2_{Y\sim W|Z,\boldsymbol{X}_{-j}}}{R^2_{Y\sim X_j|Z\boldsymbol{X}_{-j}}}. \tag{3.39}$$

where $\boldsymbol{X}_{-j}$ represents the vector of covariates $\boldsymbol{X}$ excluding X_j. These new parameters, k_Z and k_Y, stand for how much "stronger" W is relatively to the observed covariate X_j in terms of residual variation explained of Z and Y. Our goal in this section is to re-express (or bound) the sensitivity parameters $R^2_{Z\sim W|\boldsymbol{X}}$ and $R^2_{Y\sim W|Z,\boldsymbol{X}}$ in terms of the relative strength parameters k_Z and k_Y.

We start by restating the bounds derived in Section 2.4.4. These are particularly useful when contemplating X_j and W both *confounders* of Z (violations of the ignorability of the instrument). Let $R^2_{W\sim X_j|\boldsymbol{X}_{-j}} = 0$ (or, equivalently, consider the part of W not linearly explained by $\boldsymbol{X}$). Then the previous sensitivity parameters can be written as

$$R^2_{Z\sim W|\boldsymbol{X}} = k_Z f^2_{Z\sim X_j|\boldsymbol{X}_{-j}}, \qquad R^2_{Y\sim W|Z,\boldsymbol{X}} \leq \eta^2 f^2_{Y\sim X_j|Z,\boldsymbol{X}_{-j}} \tag{3.40}$$

where η is a function of both parameters k_Y, k_Z and $R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}$.

In the instrumental variable setting, however, W and X_j may be *side-effects* of Z, instead of causes of Z (violations of the exclusion restriction). In such cases, reasoning about the orthogonality of $\boldsymbol{X}$ and W may not be natural, as the instrument itself is a source of dependence between these variables. Therefore, here we additionally provide bounds under the alternative condition $R^2_{W\sim X_j|Z,\boldsymbol{X}_{-j}} = 0$. We then have that

$$R^2_{Z\sim W|\boldsymbol{X}} = \eta' f^2_{Z\sim X_j|\boldsymbol{X}_{-j}}, \qquad R^2_{Y\sim W|Z,\boldsymbol{X}} \leq k_Y f^2_{Y\sim X_j|Z,\boldsymbol{X}_{-j}} \tag{3.41}$$

where η' is a function of k_Z and $R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}$ (see appendix for details).

These results allow investigators to leverage knowledge of *relative importance* of variables [88] when making plausibility judgments regarding sensitivity parameters. For instance, if researchers have domain knowledge to argue that a certain observed covariate X_j is supposed to be a strong determinant of the instrument and the outcome variation, and that the omitted variable W is not likely to explain as much residual variance of Z and Y as that observed covariate, such results can be used to set plausible bounds on the maximum bias due to the omission of W.

3.4 An omitted variable bias framework for the sensitivity of IV

Having established the tools for analyzing the sensitivity of conventional OLS estimates, we are now in a position to develop a suite of sensitivity analysis tools for instrumental variable analyses. As explained, an OVB-approach to sensitivity begins by assuming that the researcher measured and included observed covariates $\boldsymbol{X}$, but would also have liked to adjust for W in order for the IV conditions to hold. In this section, we first show how separate sensitivity analysis of the reduced form and first stage is already sufficient to draw valuable conclusions regarding the sensitivity of IV. We then construct a complete OVB framework for sensitivity analysis of IV within the Anderson-Rubin approach, allowing one to investigate the sensitivity of tests to a specific null hypothesis, the sensitivity of lower and upper limits of confidence intervals, to define and compute sensitivity statistics for routine reporting for IV, such as (extreme) robustness values, as well as providing bounds on the sensitivity parameters, on the basis of comparison to observed covariates.

3.4.1 Sensitivity analysis of the reduced form and of the first stage

The recent literature on instrumental variables places strong emphasis on the first-stage and the reduced-form estimates. Not only are the first stage and reduced form often substantively meaningful on their own, but their critical examination plays an important role for motivating the causal story behind a particular instrumental variable. For example, in the "local average treatment effect" interpretation of the IV estimand, *both* the first stage and the reduced form must be unconfounded so that the resulting estimate can be interpreted as the average causal effect among compliers [4]. Therefore, beyond a means to the final IV estimate, researchers are advised to report and to interpret the first stage and the reduced form by, for example, assessing whether those results are in accordance to the postulated mechanisms that justify the choice of instrument [5, 6, 82, 7, 83]. While investigating these separate regressions, researchers can deploy all sensitivity analysis results discussed in the previous section.

Fortunately, such sensitivity analyses also provide answers to many pivotal sensitivity questions regarding the IV estimate itself. In particular, if the investigator is interested in assessing the strength of confounders or side-effects needed to bring the IV point estimate to zero, or to not reject the null hypothesis of zero effect, the results of the sensitivity analysis of the reduced form is all that is needed. If interest lies in also determining whether the IV estimate could be arbitrarily large in either direction, then the sensitivity of the first stage must also be assessed, as omitted variables capable of changing the direction of the first stage can lead to unbounded IV estimates. We now give a more precise meaning to these claims.

3.4.1.1 What the reduced form and first stage reveal about the IV point estimate

First let us consider the sensitivity of the point estimate. Recall that all estimators under consideration are algebraically equivalent, and are equal to the ratio of the reduced-form and the first-stage coefficients,

$$\hat{\tau} := \hat{\tau}_{\mathrm{ILS}} = \hat{\tau}_{\mathrm{2SLS}} = \hat{\tau}_{\mathrm{AR}} = \frac{\hat{\lambda}}{\hat{\theta}} \tag{3.42}$$

This simple algebraic fact allows us to draw two important conclusions regarding the sensitivity of $\hat{\tau}$ from the sensitivity of $\hat{\lambda}$ and $\hat{\theta}$ alone.

First, residual biases can bring the IV point estimate to zero *if, and only if,* they can bring the reduced-form point estimate to zero. Therefore, if sensitivity analysis of the reduced form reveals that omitted variables are not strong enough to explain away $\hat{\lambda}$, then they also cannot explain away the IV point estimate $\hat{\tau}$. Or, more worrisome, if analysis reveals that it takes weak confounding or side-effects to explain away $\hat{\lambda}$, the same holds for the IV estimate $\hat{\tau}$. In sum, for all IV estimators considered here, to assess the strength of biases needed to bring the IV point estimate to zero, one needs only to perform a sensitivity analysis on the reduced-form regression coefficient.

Second, if we cannot rule out confounders or side-effects that are sufficiently strong to *change the sign* of the first-stage point estimate $\hat{\theta}$, then we also cannot rule out that the IV point estimate $\hat{\tau}$ could be *arbitrarily large* in either direction, even if not exactly equal to zero. This can be immediately seen by letting $\hat{\theta}$ approach zero on either side of the limit. Thus, whenever we are interested in biases as large *or larger* than a certain amount, the robustness of the first stage to the zero null puts an upper bound on the robustness of the IV point estimate.

3.4.1.2 What the reduced form and first stage reveal about IV hypothesis tests

Contrary to the point estimate, the different approaches presented here may lead to different conclusions regarding how omitted variables would have changed inferences. Let us start by examining the Anderson-Rubin/Fieller approach, as not only it has nominal coverage regardless of instrument strength, but its conclusions match the intuition of current guidelines when assessing the first-stage and reduced-form estimates [5, 6, 7].

Consider again the IV estimand

$$\tau = \frac{\lambda}{\theta}$$

Note that the same arguments we used before for the estimator hold for the estimand. Logically, provided the ratio is well defined ($\theta \neq 0$), we have that $\tau = 0 \iff \lambda = 0$. Therefore, a test of the null hypothesis $H_0 : \lambda = 0$ in the reduced-form regression is *logically equivalent* to a test of the null hypothesis $H_0 : \tau = 0$ for the IV estimand. Similarly, for a fixed λ, if we cannot rule out that θ is arbitrarily close to zero in either direction, then, logically, we also cannot rule out that τ is arbitrarily large in either direction—a test for the null hypothesis $H_0 : \theta = 0$ is thus *logically equivalent* to testing whether arbitrarily large sizes for τ can be ruled out.

The Anderson-Rubin/Fieller approach is coherent with respect to these logical implications. Recall the Anderson-Rubin test for the null hypothesis $H_0 : \tau = \tau_0$ is based on the test of $H_0 : \phi_{\tau_0} = 0$. By the FWL theorem, the point estimate and (estimated) standard error for $\hat{\phi}_{\tau_0}$ are given by

$$\hat{\phi}_{\tau_0} = \frac{\text{cov}(Y_{\tau_0}^{\perp \boldsymbol{X},W}, Z^{\perp \boldsymbol{X},W})}{\text{var}(Z^{\perp \boldsymbol{X},W})}, \qquad \widehat{\text{se}}(\hat{\phi}_{\tau_0}) = \frac{\text{sd}(Y_{\tau_0}^{\perp Z,\boldsymbol{X},W})}{\text{sd}(Z^{\perp \boldsymbol{X},W})}\sqrt{\frac{1}{\text{df}-1}} \tag{3.43}$$

Which can be expressed in terms of the first-stage and reduced-form estimates (see appendix)

$$\hat{\phi}_{\tau_0} = \hat{\lambda} - \tau_0\hat{\theta}, \qquad \widehat{\text{se}}(\hat{\phi}_{\tau_0}) = \sqrt{\widehat{\text{var}}(\hat{\lambda}) + \tau_0^2\widehat{\text{var}}(\hat{\theta}) - 2\tau_0\widehat{\text{cov}}(\hat{\lambda}, \hat{\theta})} \tag{3.44}$$

Testing $H_0 : \phi_{\tau_0} = 0$ requires comparing the t-value for $\hat{\phi}_{\tau_0}$ with a critical threshold $t^*_{\alpha,\text{df}-1}$, and the null hypothesis is not rejected if $|t_{\hat{\phi}_{\tau_0}}| \leq t^*_{\alpha,\text{df}-1}$. Squaring and rearranging terms we obtain the quadratic inequality which must hold for non-rejection:

$$\underbrace{\left(\hat{\theta}^2 - \widehat{\text{var}}(\hat{\theta}) \times t^{*2}_{\alpha,\text{df}-1}\right)}_{a}\tau_0^2 + \underbrace{2\left(\widehat{\text{cov}}(\hat{\lambda}, \hat{\theta}) \times t^{*2}_{\alpha,\text{df}-1} - \hat{\lambda}\hat{\theta}\right)}_{b}\tau_0 + \underbrace{\left(\hat{\lambda}^2 - \widehat{\text{var}}(\hat{\lambda}) \times t^{*2}_{\alpha,\text{df}-1}\right)}_{c} \leq 0 \tag{3.45}$$

When considering the null hypothesis $H_0 : \tau_0 = 0$, only the term c remains, and c is less or equal to zero if, and only if, one cannot reject the null hypothesis $H_0 : \lambda = 0$ in the reduced-form regression. The Anderson-Rubin approach thus comports with the recommendation of [5] that "if you can't see the causal relation of interest in the reduced form, it's probably not there." Also note that arbitrarily large values for τ_0 will satisfy the inequality in Equation 3.45 if, and only if, $a < 0$, meaning that we cannot reject the null hypothesis $H_0 : \theta = 0$ in the first-stage regression. This supports the recommendation that, if one is unsure about the direction of the first stage, it is likely that very little can be said about the magnitude of the

IV estimate.

Within the Anderson-Rubin framework, we thus reach analogous conclusions regarding hypothesis testing as those regarding the point estimate: (i) when interest lies in the zero null hypothesis, the sensitivity of the reduced form is exactly the sensitivity of the IV—no other analyses are needed. Confounders or side-effects sufficiently strong to bring the reduced form to a region where it is not statistically different than zero can also bring the IV estimate to a region where it is not statistically different than zero, and only omitted variables with such strength are capable of doing so; and, (ii) if one is interested in biases of a certain amount, *or larger,* then the sensitivity of the first stage to the zero null hypothesis needs also to be assessed. Specifically, for any null hypothesis of interest $H_0 : \tau = \tau_0$, omitted variables that are strong enough to make the first stage not statistically different from zero may also lead us to not reject values arbitrarily "worse" than τ_0.[9]

As is well known, it is not uncommon for frequentist statistical tests to lead to logically incoherent decisions [64, 126, 105, 56]. While inferences made in the Anderson-Rubin approach have the expected behavior in this setting, inferences using ILS or 2SLS, however, do not necessarily comply with these logical expectations. Cases can be found for ILS and 2SLS where, for instance, one fails to reject the null hypothesis $H_0 : \lambda = 0$, yet still rejects the null hypothesis $H_0 : \tau = 0$ (and vice-versa). Such claims do not conform to current guidelines for interpreting the first-stage and reduced-form regressions [6].

3.4.2 Sensitivity analysis of the IV in the Anderson-Rubin approach

We now apply the OVB framework for assessing the sensitivity of the IV estimate directly. We focus on the Anderson-Rubin approach for this task because: (i) it allows performing sensitivity analysis of the IV with only two interpretable sensitivity parameters; (ii) it has correct test size regardless of "instrument strength"; and, (iii) its conclusions conform to current recommendations regarding the interpretation of the first-stage and reduced-form

[9] Similar observations regarding the importance of the robustness of the first stage for hidden biases have been made before in the context of randomization inference [131, 123].

regressions.

3.4.2.1 Sensitivity for testing a specific null hypothesis

We begin by examining the sensitivity of the t-value for testing a specific null hypothesis $H_0 : \tau = \tau_0$, as this is a straightforward application of the tools of Section 3.3. Recall that, in the Anderson-Rubin approach, a test for the null hypothesis $H_0 : \tau = \tau_0$ is a test for the null hypothesis $H_0 : \phi_{\tau_0} = 0$ in the regression of Y_{τ_0} on the instrument Z and covariates $\boldsymbol{X}$ and W. Therefore, standard OLS sensitivity analysis for testing the null hypothesis $H_0 : \phi_{\tau_0} = 0$ on the Anderson-Rubin regression gives the desired results for $H_0 : \tau = \tau_0$.

In detail, a sensitivity analysis for the null hypothesis that the IV estimate τ equals some τ_0 can be performed as follows:

1. Construct $Y_{\tau_0} = Y - \tau_0 D$ under the null value $H_0 : \tau = \tau_0$;
2. Run the OLS model $Y_{\tau_0} = \hat{\phi}_{\text{res},\tau_0} Z + \boldsymbol{X}\hat{\beta}_{\text{res},\tau_0} + \hat{\varepsilon}_{\tau_0,\text{res}}$;
3. Perform regular OLS sensitivity analysis for the null $H_0 : \phi_{\tau_0} = 0$.

This procedure can both tell us how omitted variables no worse than $\mathbf{R}^2 = \{R^2_{Z\sim W|\boldsymbol{X}}, R^2_{Y_{\tau_0}\sim W|Z,\boldsymbol{X}}\}$ would alter inferences regarding the null $H_0 : \tau = \tau_0$, or what is the minimal strength of $\mathbf{R}^2$ that is required to not reject the null $H_0 : \tau = \tau_0$, as given by the RV or XRV.

Making sense of the sensitivity parameters. While separate analyses of the first stage and reduced form regressions may suggest the need of three sensitivity parameters for the sensitivity of IV (e.g, $R^2_{Z\sim W|\boldsymbol{X}}$, $R^2_{D\sim W|Z,\boldsymbol{X}}$ and $R^2_{Y\sim W|Z,\boldsymbol{X}}$), note how within the Anderson-Rubin approach one is able to perform sensitivity with only two parameters ($R^2_{Z\sim W|\boldsymbol{X}}, R^2_{Y_{\tau_0}\sim W|Z,\boldsymbol{X}}$). The meaning of the parameter related with the instrument ($R^2_{Z\sim W|\boldsymbol{X}}$) is unchanged and straightforward, ie., the share of residual variation of the instrument explained by the omitted variable W. The main difference concerns the parameter $R^2_{Y_{\tau_0}\sim W|Z,\boldsymbol{X}}$, which stands for the share of residual variance of Y_{τ_0} explained by W. The substantive interpretation of Y_{τ_0} depends on the causal assumptions the researcher is willing to defend. For instance, under $H_0 : \tau = \tau_0$ and a constant treatment effects model, we have that $Y_{\tau_0} = Y - \tau_0 D$ equals the *untreated potential outcome* Y_0 and thus $R^2_{Y_{\tau_0} \sim W|Z,\boldsymbol{X}}$ could be interpreted as the share of residual variance of Y_0 explained by W. For simplicity of exposition, we adopt this interpretation throughout the chapter.

3.4.2.2 Compatible inferences given bounds on partial R^2

Instead of assessing the sensitivity of the test statistic for specific a null hypothesis, investigators may be interested in recovering the whole set of inferences compatible with plausibility judgments on the maximum strength of W. As discussed in Section 3.2, for a critical threshold $t^*_{\alpha,\mathrm{df}-1}$, the confidence interval for τ in the Anderson-Rubin framework is given by

$$\mathrm{CI}_{1-\alpha}(\tau) = \{\tau_0;\ t^2_{\phi_{\tau_0}} \leq t^{*2}_{\alpha,\mathrm{df}-1}\} \tag{3.46}$$

Now consider bounds on sensitivity parameters $R^2_{Y_{\tau_0} \sim W|Z,\boldsymbol{X}} \leq R^{2\,\max}_{Y_0 \sim W|Z,\boldsymbol{X}}$ (which should be judged to hold *regardless* of the value of τ_0) and $R^2_{Z \sim W|\boldsymbol{X}} \leq R^{2\,\max}_{Z \sim W|\boldsymbol{X}}$. Let $t^{\dagger\,\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ denote the maximum OVB-adjusted critical value under the posited bounds on the strength of W. The set of compatible inferences for τ, $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau)$ is then simply given by

$$\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau) = \left\{\tau_0;\ t^2_{\hat{\phi}_{\mathrm{res},\tau_0}} \leq \left(t^{\dagger\,\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}\right)^2\right\} \tag{3.47}$$

This interval can be found analytically using the same inequality as in Equation 3.45, now with the parameters of the restricted regression actually run, and the traditional critical value replaced by the OVB-adjusted critical value $t^{\dagger\,\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$

$$\begin{aligned}\underbrace{\left(\hat{\theta}^2_{\mathrm{res}} - \widehat{\mathrm{var}}(\hat{\theta}_{\mathrm{res}}) \times \left(t^{\dagger\,\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}\right)^2\right)}_{a}\tau_0^2 + 2\underbrace{\left(\widehat{\mathrm{cov}}(\hat{\lambda}_{\mathrm{res}}, \hat{\theta}_{\mathrm{res}}) \times \left(t^{\dagger\,\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}\right)^2 - \hat{\lambda}_{\mathrm{res}}\hat{\theta}_{\mathrm{res}}\right)}_{b}\tau_0 \\ + \underbrace{\left(\hat{\lambda}^2_{\mathrm{res}} - \widehat{\mathrm{var}}(\hat{\lambda}_{\mathrm{res}}) \times \left(t^{\dagger\,\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}\right)^2\right)}_{c} \leq 0\end{aligned} \tag{3.48}$$

Note that users can easily obtain $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau)$ with any software that computes Anderson-Rubin or Fieller's confidence intervals by simply providing the modified critical threshold $t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$.

It is now useful to discuss the possible shapes of $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}$ as this will help understanding the robustness values for IV we derive next. Let $\mathbf{r} = \{r_{\min}, r_{\max}\}$ denote the roots of the quadratic equation, which can be written as $\mathbf{r} = -b \pm \sqrt{\Delta}/2a$, with $\Delta = b^2 - 4ac$. If $a > 0$ (i.e, we have a statistically significant first stage), the quadratic equation will be convex, and thus only the values between the roots will be non-positive. This leads to the connected confidence interval $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2} = [r_{\min}, r_{\max}]$. When $a < 0$ (i.e, the null hypothesis of zero for the first stage is not rejected), the curve is concave and this leads to unbounded confidence intervals. Here we have two sub-cases: (i) when $\Delta < 0$, the quadratic curve never touches zero, and thus the confidence interval is simply the whole real line $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2} = (-\infty, +\infty)$; and, (ii) when $\Delta > 0$ the confidence interval will be union of two disjoint intervals $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2} = (-\infty, r_{\min}] \cup [r_{\max}, +\infty)$.[10]

3.4.2.3 Sensitivity statistics for routine reporting

Armed with the notion of a set of compatible inferences for IV, $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau)$, we are now able to formally define and derive (extreme) robustness values for instrumental variable estimates.

Extreme robustness values for IV. The extreme robustness value $\mathrm{XRV}_{q^*,\alpha}(\tau)$ for the IV estimate is defined as the minimum strength of association of omitted variables with the instrument so that we cannot reject a reduction of $(100 \times q^*)\%$ of the original IV estimate; that is,

$$\mathrm{XRV}_{q^*,\alpha}(\tau) := \inf\left\{\mathrm{XRV};\ (1-q^*)\hat{\tau}_{\mathrm{res}} \in \mathrm{CI}^{\max}_{1-\alpha,1,\mathrm{XRV}}(\tau)\right\} \tag{3.49}$$

[10] See [98] for an intuitive graphical characterization of Fieller's solutions using polar coordinates.

It then follows immediately from Equation 3.47 that

$$\mathrm{XRV}_{q^*,\alpha}(\tau) = \mathrm{XRV}_{1,\alpha}(\phi_{\tau^*}) \tag{3.50}$$

where $\tau^* = (1-q^*)\hat{\tau}_{\text{res}}$. As in the general case, the extreme robustness value can be interpreted as a "dampened" partial R^2 of the instrument Z with the "putative" untreated potential outcome Y_{τ_0}. Also of interest is the special case of the minimum strength to bring the IV estimate to a region where it is no longer statistically different than zero ($q^* = 1$), in which we obtain $\mathrm{XRV}_{1,\alpha}(\tau) = \mathrm{XRV}_{1,\alpha}(\lambda)$. That is, for the null hypothesis of $H_0 : \tau = 0$, the extreme robustness value of the IV estimate equals the extreme robustness value of the reduced-form estimate, as we discussed in the last section.

The $\mathrm{XRV}_{q^*,\alpha}(\tau)$ computes the minimal strength of W required to not reject a particular null hypothesis of interest. We might be interested, instead, in asking about the minimal strength of omitted variables to not reject a specific value *or worse*. When confidence intervals are connected, such as the case of standard OLS, the two notions coincide. But in the Anderson-Rubin case, as we have seen, confidence intervals for the IV estimate can sometimes consist of disjoint intervals. Therefore, let the upper and lower limits of $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R^2}}(\tau)$ be $\mathrm{LL}^{\max}_{1-\alpha,\boldsymbol{R^2}}(\tau)$ and $\mathrm{UL}^{\max}_{1-\alpha,\boldsymbol{R^2}}(\tau)$ respectively. The extreme robustness value $\mathrm{XRV}_{\geq q^*,\alpha}(\tau)$ for the IV estimate is defined as the minimum strength of association that confounders or side-effects need to have with the instrument so that we cannot reject a change of $(100 \times q^*)\%$ *or worse* of the original IV estimate;

$$\mathrm{XRV}_{\geq q^*,\alpha}(\tau) := \inf\left\{\mathrm{XRV};\ (1-q^*)\hat{\tau}_{\text{res}} \in \left[\mathrm{LL}^{\max}_{1-\alpha,1,\mathrm{XRV}}(\tau),\quad \mathrm{UL}^{\max}_{1-\alpha,1,\mathrm{XRV}}(\tau)\right]\right\} \tag{3.51}$$

Now note that, whenever $\mathrm{CI}^{\max}_{1-\alpha,\mathrm{df}-1}(\tau)$ is connected, we must have that $\mathrm{XRV}_{\geq q^*,\alpha}(\tau) = \mathrm{XRV}_{q^*,\alpha}(\tau)$. On the other hand, recall that $\mathrm{CI}^{\max}_{1-\alpha,\mathrm{df}-1}(\tau)$ will be disjoint only if $t^2_{\hat{\theta}_{\text{res}}} \leq (t^{\dagger\max}_{\alpha,\mathrm{df}-1})^2$, which is precisely the condition for the extreme robustness value of the first stage.

Therefore,

$$\text{XRV}_{\geq q^*,\alpha}(\tau) = \min\{\text{XRV}_{1,\alpha}(\phi_{\tau^*}), \quad \text{XRV}_{1,\alpha}(\theta)\} \tag{3.52}$$

This corroborates our previous conclusion that, when we are interested in biases as large or larger than a certain amount, the robustness of the IV estimate is bounded by the robustness of the first stage assessed at the zero null.

Robustness values for IV. The definitions of the robustness value for IV follow the same logic discussed above, but now considering both bounds on $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}$ varying simultaneously. That is,

$$\text{RV}_{q^*,\alpha}(\tau) := \inf\left\{\text{RV};\ (1-q^*)\hat{\tau}_{\text{res}} \in \text{CI}^{\max}_{1-\alpha,\text{RV},\text{RV}}(\tau)\right\} \tag{3.53}$$

Again from Equation 3.47 we have that

$$\text{RV}_{q^*,\alpha}(\tau) = \text{RV}_{1,\alpha}(\phi_{\tau^*}) \tag{3.54}$$

Which for the special case of $q^* = 1$ simplifies to $\text{RV}_{1,\alpha}(\tau) = \text{RV}_{1,\alpha}(\lambda)$, as before. We can also define robustness values for not rejecting the null hypothesis of a reduction of $(100 \times q^*)\%$ *or worse*

$$\text{RV}_{\geq q^*,\alpha}(\tau) := \inf\left\{\text{RV};\ (1-q^*)\hat{\tau}_{\text{res}} \in \left[\text{LL}^{\max}_{1-\alpha,\text{RV},\text{RV}}(\tau), \quad \text{UL}^{\max}_{1-\alpha,\text{RV},\text{RV}}(\tau)\right]\right\} \tag{3.55}$$

By the same arguments articulated above, $\text{RV}_{\geq q^*,\alpha}(\tau)$ must be the minimum of the robustness value of the Anderson-Rubin regression evaluated at $\tau^* = (1-q^*)\hat{\tau}_{\text{res}}$ and the robustness value of the first-stage regression evaluted at the zero null

$$\text{RV}_{\geq q^*,\alpha}(\tau) = \min\{\text{RV}_{1,\alpha}(\phi_{\tau^*}), \quad \text{RV}_{1,\alpha}(\theta)\} \tag{3.56}$$

For the special case of $q^* = 1$ (zero null hypothesis), $\text{RV}_{\geq q^*,\alpha}(\tau)$ simplifies to the minimum of the robustness value of the first stage and of the reduced form, $\text{RV}_{\geq q^*=1,\alpha}(\tau) = \min\{\text{RV}_{1,\alpha}(\lambda), \quad \text{RV}_{1,\alpha}(\theta)\}$.

3.4.2.4 Bounds on the strength of omitted variables

The bounds discussed in Section 3.3.3 work without any major modifications in the Anderson-Rubin setting. When testing a specific null hypothesis $H_0 : \tau = \tau_0$ in the AR regression, we have k_Z as before, and instead of k_Y we now have $k_{Y_{\tau_0}}$

$$k_{Y_{\tau_0}} := \frac{R^2_{Y_{\tau_0}\sim W|Z,\boldsymbol{X}_{-j}}}{R^2_{Y_{\tau_0}\sim X_j|Z\boldsymbol{X}_{-j}}}. \tag{3.57}$$

The plausibility judgment one is making here is that of how strong unobserved confounders or side-effects are, relative to observed covariates, in explaining the residual variance of the untreated potential outcome and of the instrument, *under the null hypothesis* $H_0 : \tau = \tau_0$.

Since the judgment is made under a specific null, the bounds will be different when testing different hypotheses. Therefore, it may be useful to compute bounds under a slightly more *conservative* assumption. More precisely, consider

$$k^{\max}_{Y_{\tau_0}} := \frac{\max_{\tau_0} R^2_{Y_{\tau_0}\sim W|Z,\boldsymbol{X}_{-j}}}{\max_{\tau_0} R^2_{Y_{\tau_0}\sim X_j|Z\boldsymbol{X}_{-j}}}. \tag{3.58}$$

That is, we can posit that the omitted variables are no stronger than (a multiple of) the *maximum* explanatory power of an observed covariate, regardless of the value of τ_0. This has the useful property of providing a unique valid bound for any value of the null hypothesis, and can be used to place bounds on sensitivity contours of the lower and upper limit of the AR confidence intervals, as we show next.

3.5 Using the OVB framework for the sensitivity analysis of IV

In this section we return to our running example of estimating the returns to schooling and show how these tools can be deployed to assess the robustness of those findings to violations of the IV assumptions. We propose investigators begin their sensitivity analysis by examining the robustness of the first-stage and reduced-form estimates. Not only are these analyses usually important on their own right, but in many cases—including this one—this exercise will be sufficient to establish that the instrumental variable estimate is not very informative of the causal effect of interest, since one is not in a position to rule out confounders or side-effects that can explain away those auxiliary estimates. We then turn to the sensitivity of the IV itself, and further show how sensitivity contour plots of the adjusted lower and upper limits of the AR confidence interval, supplemented with benchmark bounds, give a succinct yet complete picture of the whole range of sensitivity of the IV estimate.

3.5.1 Minimal reporting and sensitivity plots of the reduced form

Outcome: *Earnings* (log)

Instrument	Estimate	Std. Error	t-value	$R^2_{Y\sim Z\|\boldsymbol{X}}$	$\text{XRV}_{q^*,\alpha}$	$\text{RV}_{q^*,\alpha}$
Proximity	0.042	0.018	2.33	0.18%	0.05%	0.67%

Bound (1x SMSA): $R^2_{Y\sim W|Z,\boldsymbol{X}} = 2\%$, $R^2_{W\sim Z|\boldsymbol{X}} = 0.6\%$, $t^{\dagger\max}_{\alpha,\text{df}-1,\boldsymbol{R}^2} = 2.55$

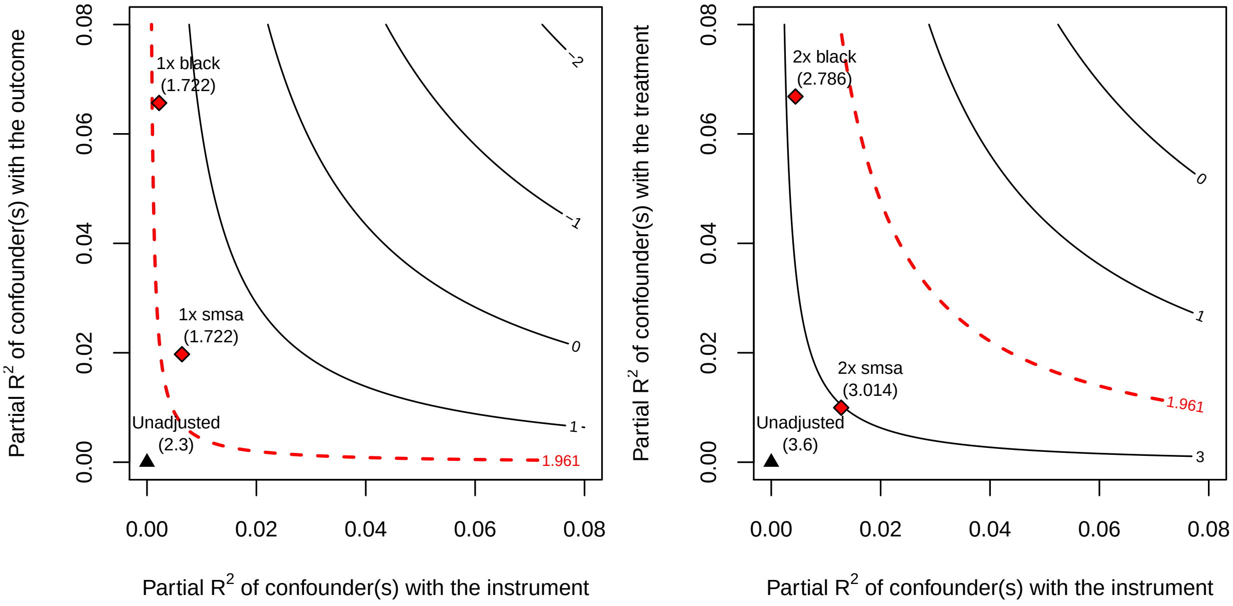
Note: df $= 2994$, $q^* = 1$, $\alpha = 0.05$

Table 3.3: Minimal sensitivity reporting of the reduced-form regression.

We start by examining the sensitivity of the reduced-form estimate, namely, the effect of *Proximity* on *Earnings*. Recall that if we cannot rule out that the reduced form is zero, we also cannot rule out the IV estimate is zero. In our running example we focus the discussion on violations of the ignorability of the instrument due to confounders, as this is the main threat of the study under investigation. Readers should keep in mind, however, that all analyses performed here can be equally used to assess violations of the exclusion restriction (or both). Table 3.3 shows the minimal sensitivity reporting we proposed in Section 2.5.1, but now incorporating the new results of Section 3.3. Beyond the usual statistics such as the point estimate, standard-error and t-value, we recommend that researchers also report the: (i) partial R^2 of the instrument with the outcome ($R^2_{Y\sim Z|\boldsymbol{X}} = 0.18\%$), as well as (ii) the robustness value ($\mathrm{RV}_{q^*,\alpha} = 0.67\%$), and (iii) the extreme robustness value ($\mathrm{XRV}_{q^*,\alpha} = 0.05\%$), both for where the confidence interval would cross zero ($q^* = 1$), at a chosen significance level (here, $\alpha = 0.05$).

For our running example, the robustness value reveals that confounders that explain 0.67% of the residual variation both of *proximity* and of (log) *Earnings* are sufficiently strong to make the reduced-form estimate statistically insignificant, whereas confounders that explain less than 0.67% of the residual variation of both the instrument and of the outcome are not strong enough to do so. The extreme robustness value and the partial R^2 show that, if we are not willing to impose constraints on the strength of confounders with the outcome, then they would need to explain less than 0.05% or 0.18% of the instrument to escape concerns of eliminating statistical significance or fully eliminating the point estimate, respectively. To aid users in making plausibility judgments, the note of Table 3.3 provides the maximum strength of unobserved confounding if it were as strong as *SMSA* (an indicator variable for whether the individual lived in a metropolitan region) along with the OVB-adjusted critical value for a confounder with such strength, $t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} = 2.55$. Since the observed t-value (2.33) is less than the adjusted critical threshold of 2.55, the table immediately reveals that confounding as strong as *SMSA* (for example, in the form of residual geographic confounding) is sufficiently strong to be problematic.

Beyond the results of Table 3.3, we also advise researchers to provide a sensitivity contour plot of the t-value for testing the null hypothesis of zero effect, while also showing different bounds on strength of confounding, under different assumptions of how they compare to the observed variables. This is shown in Figure 3.3a. The horizontal axis describes the partial R^2 of the confounder with the instrument whereas the vertical axis describes the partial R^2 of the confounder with the outcome. The contour lines show the t-value one would have obtained, had a confounder with such postulated strength been included in the reduced-form regression. The red dashed line shows the statistical significance threshold, and the red



(a) Sensitivity contours of the reduced form.

(b) Sensitivity contours of the first stage.

Figure 3.3: Sensitivity contour plots with benchmark bounds for the t-value of: (a) the reduced form; and, (b) the first stage.

diamonds places bounds on strength of confounding as strong as *Black* (an indicator for race) and, again, *SMSA*. As we can see, confounders as strong as either *Black* or *SMSA* are sufficient to bring the reduced form, and hence also the IV estimate, to a region which is not statistically different from zero. Since it is not very difficult to imagine residual confounders as strong or stronger than those (e.g., parental income, finer grained geographic location, etc), these results for the reduced form are sufficient to call into question the reliability of the instrumental variable estimate.

3.5.2 Minimal reporting and sensitivity plots of the first stage

Outcome: *Education* (years)

Instrument	Estimate	Std. Error	t-value	$R^2_{D\sim Z\|\boldsymbol{X}}$	$\text{XRV}_{q^*,\alpha}$	$\text{RV}_{q^*,\alpha}$
Proximity	0.32	0.088	3.64	0.44%	0.31%	3.02%

Bound (1x SMSA): $R^2_{D\sim W|Z,\boldsymbol{X}} = 0.5\%$, $R^2_{Z\sim W|\boldsymbol{X}} = 0.6\%$, $t^{\dagger\,\max}_{\alpha,\text{df}-1,\boldsymbol{R}^2} = 2.26$

Note: df $= 2994$, $q^* = 1$, $\alpha = 0.05$

Table 3.4: Minimal sensitivity reporting of the first-stage regression.

We now turn to the sensitivity analysis of the first-stage regression. Table 3.4 performs the same sensitivity exercises as before, but now for the regression of *Education* (treatment) on *Proximity* (instrument). As expected, the association of proximity to college with years of education is stronger than its association with earnings, and this is also reflected in the robustness statistics, which are slightly higher ($R^2_{D\sim Z|\boldsymbol{X}} = 0.44\%$, $\text{XRV}_{q^*,\alpha} = 0.31\%$ and $\text{RV}_{q^*,\alpha} = 3.02\%$). As the note of Table 3.4 shows, confounding as strong as *SMSA* would not be sufficiently strong to bring the first-stage estimate to a region where it is not statistically different than zero. Figure 3.3b supplements those analysis with the sensitivity contour plot for the t-value of the first-stage regression. Here the horizontal axis still describes the partial R^2 of the confounder with the instrument, but now the vertical axis describes the partial R^2 of the confounder with the treatment. The plot reveals that, contrary to the reduced form, the first stage survives confounding once or twice as strong as *Black* or *SMSA*. The contrast of both sensitivity results suggests that, in our running example, the most prominent risk to the validity of the IV estimate comes from residual confounding on the reduced-form estimate.

3.5.3 Minimal reporting and sensitivity plots of the IV

Outcome: *Earnings* (log)

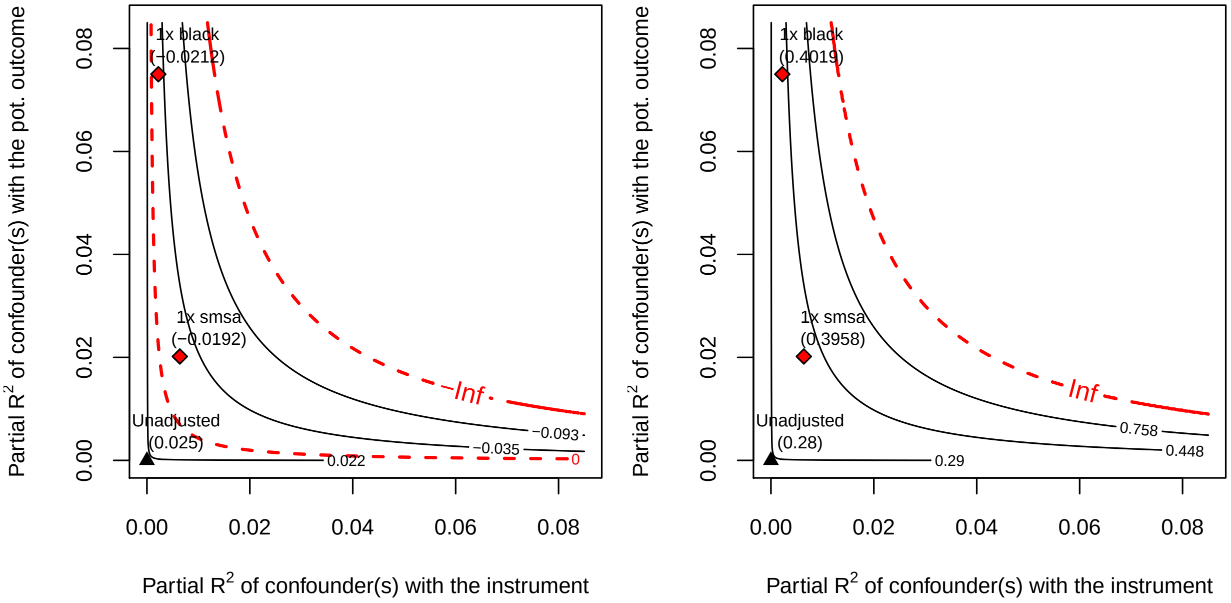
Treatment	Estimate	$\text{LL}_{1-\alpha}$	$\text{UL}_{1-\alpha}$	t-value	$\text{XRV}_{\geq q^*,\alpha}$	$\text{RV}_{\geq q^*,\alpha}$
Education (years)	0.132	0.025	0.285	2.33	0.05%	0.67%

Bound (1x SMSA): $R^2_{Y_0\sim W|Z,\boldsymbol{X}} = 2\%$, $R^2_{W\sim Z|\boldsymbol{X}} = 0.6\%$, $t^{\dagger\,\max}_{\alpha,\text{df}-1,\boldsymbol{R}^2} = 2.55$

Note: df = 2994, $q^* = 1$, $\alpha = 0.05$

Table 3.5: Minimal sensitivity reporting of IV estimate (Anderson-Rubin).

Finally, we turn our attention to the sensitivity analysis of the IV, and Table 3.5 shows our proposed minimal sensitivity reporting. We start with the IV point estimate (0.132), as well as the lower limit ($\text{LL}_{1-\alpha} = 0.025$) and the upper limit ($\text{UL}_{1-\alpha} = 0.285$) of the Anderson-Rubin confidence interval. The t-value for testing the null hypothesis of zero effect is also shown (2.33). Next, we propose researchers to report the extreme robustness value $\text{XRV}_{\geq q^*,\alpha}$ and the robustness value $\text{RV}_{\geq q^*,\alpha}$ for bringing the lower limit of the confidence interval to *or beyond* zero (or another meaningful threshold), at the 5% significance level. As derived in Section 3.4.2.3, we have that the (extreme) robustness value of the IV estimate for bringing the lower limit of the confidence interval to or below zero is the minimum of either the (extreme) robustness value of the reduced form and the (extreme) robustness value of the first stage. Therefore, the sensitivity statistics of Table 3.5 essentially reproduce the results of Table 3.3.



(a) Sensitivity contours for the lower limit.

(b) Sensitivity contours for the upper limit.

Figure 3.4: Sensitivity contour plots for the lower (a) and upper (b) limits of the 95% confidence interval for the IV estimate.

After examining the sensitivity of the first stage and reduced form it is thus more informative to assess the sensitivity of the IV for null hypotheses *other than zero.* To that end, investigators may wish to examine sensitivity contour plots similar to those of Figure 3.3, but with contours now showing the adjusted *lower and upper limits* of the confidence interval. These contours are shown Figure 3.4. Here, as usual, the horizontal axis describes the partial R^2 of the confounder with the instrument, but now the vertical axis describes the partial R^2 of the confounder with the untreated *potential* outcome. The contour lines show the worst lower (or upper) limit of the set of compatible inferences considering confounders bounded by such strength. Red dashed lines shows a critical contour line of interest (such as zero) as well as the boundary beyond confidence intervals become unbounded. As the plot reveals, even confounding as strong as *SMSA* could lead to an interval of compatible inferences for the causal effect of $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau) = [-0.02, 0.40]$, which includes not only the original OLS estimate (7.5%), but also implausibly high values (40%), or even negative values (-2%), and is thus too wide for any meaningful conclusions regarding the "true" returns to schooling. That is, if we are concerned that omitted variables as strong as *SMSA* might exist, then we are unable to reject any estimates in this range, calling into question the strength of evidence provided by this IV study.

3.6 Conclusion

In this chapter we developed a suite of sensitivity analysis tools for IV that naturally handles multiple "side-effects" and confounders of the instrument, does not require assumptions on the functional form of such omitted variables, and allows exploiting expert knowledge to bound sensitivity parameters. In particular, we introduced new sensitivity statistics for IV estimates that are suited for routine reporting, such as (extreme) robustness values, describing the minimum strength that omitted variables need to have, both with the instrument, and with the untreated potential outcome, to overturn the conclusions of an IV study. We also introduced a novel "OVB-adjusted" critical value that allows researchers to easily perform hypothesis tests or construct confidence intervals that accounts for omitted variable bias of any postulated strength, by simply replacing traditional critical values with the adjusted ones. Finally, we showed how intuitive visual displays can be deployed to fully characterize the sensitivity of IV to violations of its standard assumptions. Extension of these sensitivity analysis tools beyond the "just-identified" case is an interesting direction for future work.

CHAPTER 4

Sensitivity Analysis of Linear Structural Causal Models

4.1 Introduction

Randomized controlled trials (RCT) are considered the gold standard for identifying cause-effect relationships in data-intensive sciences [69]. In practice, however, direct randomization is often infeasible or unethical, requiring researchers to combine non-experimental observations with assumptions about the data generating process in order to obtain causal claims. These assumptions are usually encoded as the absence of certain causal relationships, or as the absence of association between certain unobserved factors. Conclusions based on causal models are, therefore, provisional: they depend on the validity of causal assumptions, regardless of the sample size [109, 133].

In many real settings, it is not uncommon that these assumptions are subject to uncertainty or dispute. Scientists may posit alternative causal models that are equally compatible with the observed data; or, more mundanely, researchers can make identification assumptions for convenience, simply to proceed with estimation.[1] Regardless of the motivation, the provisional character of causal inference behooves us to formally assess the extent to which causal conclusions are *sensitive* to violations of those assumptions.

The importance of such exercises is best illustrated with a real example, which directly impacted public policy. During the late 1950s and early 1960s, there was a fierce debate regarding the causal effect of cigarette smoking on lung cancer. One of its most notable skeptics was the influential statistician Ronald Fisher, who claimed that, without an experiment,

[1] As noted by [86], "such assumptions are usually made casually, largely because they justify the use of available statistical methods and not because they are truly believed."

one cannot rule out unobserved common causes (e.g. the individual's genotype) as being responsible for the observed association [53, 54]. Technically speaking, Fisher's statement was accurate; data alone could not refute his hypothesis. Yet, although no RCT measuring the effect of cigarette smoking on lung cancer was performed, currently there exists a broad consensus around the issue. How could such a consensus emerge?

An important step towards the current state of affairs was a *sensitivity analysis* performed by [42]. Their investigation consisted of the following *hypothetical* question: if Fisher's hypothesis were true, *how strong* would the alleged confounder need to be *to explain all* the observed association between cigarette smoking and lung cancer? The analysis concluded that, since smokers had nine times the risk of nonsmokers for developing lung cancer, the latent confounder would need to be at least nine times more common in smokers than in nonsmokers—something deemed implausible by experts at the time.

Cornfield's exercise reveals the fundamental steps of a sensitivity analysis. The analyst introduces a *violation* of a causal assumption of the current model, such as positing the presence of unobserved confounders that induce a non-zero association between two error terms. Crucially, however, we are willing to tolerate this violation up to a certain *plausibility limit* dictated by expert judgment (e.g., prior biological understanding, pilot studies). The task is, thus, to systematically quantify how different hypothetical "degrees" of violation (to be defined) affect the conclusions, and to judge whether expert knowledge can rule out problematic values.

The problem of sensitivity analysis has been studied throughout the sciences, ranging from statistics [124, 130, 122, 35, 62] to epidemiology [21, 141, 48, 9], sociology [58], psychology [97], political science [80, 17], and economics [90, 81, 104, 96]. Notwithstanding all this attention, the current literature is still limited to specific model structures and solved on a case-by-case basis. As current practices produce a steady stream of published results, it is important to handcraft sensitivity analysis tools for widely used models, such as what we have done in Chapters 2 and 3 for identification via covariate adjustment and instrumental variable regression. However, moving forward, a formal *algorithmic* framework to deal with violations of causal assumptions is needed.

Causal modeling requires a formal language where the characterization of the data generating process can be encoded explicitly. Structural Causal Models (SCMs) [109] provide such a language and, in many fields, including machine learning, the health and social sciences, linearity is a popular modeling choice. In this chapter, we focus on the sensitivity analysis of linear acyclic semi-Markovian SCMs. We allow violations of exclusion and independence restrictions, such as (i) the absence or presence of unobserved common causes; and, (ii) the absence, presence or reversal of direct causal effects. Our contributions are the following:

1. We introduce a formal, algorithmic approach for sensitivity analysis in linear SCMs and show it can be reduced to a problem of *identification with non-zero constraints*, i.e, identification when certain parameter values are fixed to a known, but non-zero, number.

2. We develop a novel graphical procedure, called PushForward, that reduces identification with a known error covariance to vanilla identification, for which a plethora of algorithms are available.

3. We develop an efficient graph-based constrained identification algorithm that takes as input a set of sensitivity parameters and returns a sensitivity curve for the effect estimate. The algorithm is theoretically sound and experimental results corroborate its generality, showing canonical sensitivity analysis examples are a small subset of the cases solved by our proposal (within the class of linear SCMs).

This chapter is structured as follows. Section 4.2 reviews basic terminology and definitions that will be used throughout the chapter. Section 4.3 shows how sensitivity analysis in the context of linear SCMs can be reduced to a constrained identification problem. In Section 4.4 we develop a novel approach that allows researchers to systematically incorporate constraints on error covariances of linear SCMs. Section 4.5 utilizes these results to construct a constrained identification algorithm for deriving sensitivity curves. Finally, Section 4.6 presents experimental results to evaluate our proposals.

4.2 Preliminaries

In this chapter, we use the language of structural causal models as our basic semantic framework [109]. In particular, we consider linear semi-Markovian SCMs, consisting of a set of equations of the form $V = \Lambda V + U$, where V represent the endogenous variables, U the exogenous variables, and Λ a matrix containing the *structural coefficients* representing both the strength of causal relationships and lack of direct causation among variables (when $\lambda_{ij} = 0$). The exogenous variables are usually assumed to be multivariate Gaussian with covariance matrix $\mathcal{E}$, encoding independence between error terms (when $\varepsilon_{ij} = 0$).[2] We focus on acyclic models, where Λ can be arranged to be lower triangular.

The covariance matrix Σ of the endogenous variables induced by model M is given by $\Sigma = (I - \Lambda)^{-1}\mathcal{E}(I - \Lambda)^{-\top}$. Without loss of generality, we assume model variables have been standardized to unit variance. For any three variables x, y and z, we denote σ_{yx} to be the covariance of x and y, $\sigma_{yx.z}$ to be the partial covariance of y and x given z, and $R_{yx.z}$ the regression coefficient of y on x adjusting for z. Causal quantities of interest in a linear SCM are usually entries of Λ (or functions of those entries), and identifiability reduces to checking whether they can be uniquely computed from the observed covariance matrix Σ.

Causal graphs provide a parsimonious encoding of some of the substantive assumptions of a linear SCM. The causal graph (or the path diagram) of model M is a graph $G = (V, D, B)$, where V denotes the vertices (endogenous variables), D the set of directed edges (non-zero entries of Λ) and B the set of bidirected edges (non-zero entries of $\mathcal{E}$). Missing directed edges represent *exclusion restrictions*—a variable is not a direct cause of the other. Missing bidirected edges denote *independence restrictions*, representing the fact that no latent common causes exist between two observed variables. When clear from context, we may treat model coefficients and their corresponding edges on the graph interchangeably. We use standard graph notation, where $Pa(y)$ denotes the parents, $Ch(y)$ the children, $Anc(y)$ the ancestors, and $De(y)$ the descendants of node y.

[2] Gaussianity is not necessary for the results of this chapter.

4.3 Sensitivity analysis and identification

In this section we demonstrate the pervasiveness of identification problems in sensitivity analysis in the context of a simple example. Suppose a scientist hypothesizes model G_O shown in Figure 4.1a with the goal of estimating the direct effect of a treatment x on an outcome y (structural coefficient λ_{xy}). By the single-door criterion (Pearl 2000), she verifies λ_{xy} is identifiable in G_O and equal to the regression estimand $R_{yx.z}$, licensing her to proceed with estimation.

Another investigator, however, is suspicious of the bold assumption that no common causes (confounders) exist between z and x in G_O. She goes on, therefore, and constructs an alternative model G_A (Figure 4.1b) such that the bidirected edge $z \leftrightarrow x$ is included to account for that possibility. A question now naturally arises: how wrong could one be using $R_{yx.z}$ to estimate λ_{xy} if the true causal model were given by graph G_A? Answering this question requires defining a measure of "wrongness" of the estimand, and perhaps the simplest such measure is its *bias* in the additive scale.[3]

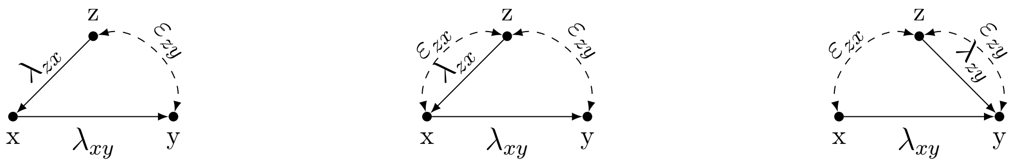
Definition 1 (Bias of ES with respect to Q)**.** *Let Q be a computable quantity given a* fully specified *linear structural causal model, and let ES be any estimand (a functional of the covariance matrix Σ). The* bias *of ES with respect to Q is the difference between the two quantities, $B = ES - Q$.*

In our example, the proposed estimand is $ES = R_{yx.z}$, the target quantity is $Q = \lambda_{xy}$, and to compute the bias, $B = R_{yx.z} - \lambda_{xy}$, one needs to *identify* λ_{xy}. Computing the bias, thus, entails an identification problem (Proposition 1).

Proposition 1. *The bias of estimand ES with respect to target quantity Q is identifiable iff Q is identifiable.*

In G_A, however, the presence of the bidirected edge $x \leftrightarrow z$ renders λ_{xy} *unidentifiable*, and computation of B is not possible. How could one circumvent this impediment?

[3] Note this refers to the bias of an *estimand* (*not* an estimator), and it is the difference between the proposed estimand and the desired (causal) target quantity in the *population*.



(a) Model G_O

(b) Model G_A

(c) Model G_B

Figure 4.1: Original model G_O and two alternative models, G_A and G_B. In G_A any of the remaining parameters (λ_{zx}, ε_{zx} or ε_{zy}) can be used as a sensitivity parameter for λ_{xy}, whereas G_B rules out ε_{zx} as a sensitivity parameter. Adding a bidirected edge $x \leftrightarrow y$ in G_A does not prevent ε_{zy} from being a valid sensitivity parameter, whereas in G_B it does.

As in [42], the impossibility of computing the exact bias of $R_{yx.z}$ with respect to λ_{xy} calls for another strategy—expressing the bias as a function of the "strength" of the omitted confounders. In this way, the analyst can predict for any *hypothetical* strength of the confounders whether it would be enough to change the research conclusions. This allows the analyst to bring new substantive knowledge to bear, by submitting these quantitative results to a judgment of plausibility and ruling out some scenarios.

Implementing this idea requires a precise definition of how to measure the "strength" of the omitted confounders. In our example, a possible candidate for measuring such strength is the structural parameter ε_{zx} of the added bidirected edge $z \leftrightarrow x$. The task then becomes: (i) to determine whether knowledge of ε_{zx} allows the identification of λ_{xy}; and, (ii) if so, to find a parameterized estimand for λ_{xy} in terms of ε_{zx}. This 2-step procedure can be seen as an identification problem with non-zero constraints (Definition 2).[4]

Definition 2 (θ-identifiability)**.** *Let M be a linear SCM and θ a set of parameters of M with known (non-zero) values. A causal quantity Q is said to be θ-identifiable if Q is uniquely computable from Σ and θ.*

We call any functional of Σ and θ, which *identifies Q*, a *θ-specific estimand* (or sensitivity curve) for Q with respect to *sensitivity parameters θ*. These estimands are the workhorse for sensitivity analysis; they allow us to investigate how strong certain relationships must be

[4] Note the relationship to z-ID [15], in which case constraints are imposed on experimental distributions in the non-parametric setting.

(as parameterized by θ) in order to induce significant bias in our estimates. In other words, identifying a bias function in terms of θ (and the observed data) for sensitivity analysis is equivalent to the constrained identification problem of Definition 2 (Proposition 2).

Proposition 2. *The bias of ES with respect to Q can be expressed as a function of θ (and Σ) iff Q is θ-identifiable.*

Going back to G_A, it is indeed possible to construct an ε_{zx}-specific estimand for λ_{xy} (see Section 4.4):

$$\lambda_{xy}(\varepsilon_{zx}) = \frac{\sigma_{xy} - (\sigma_{zx} - \varepsilon_{zx})\sigma_{yz}}{1 - (\sigma_{zx} - \varepsilon_{zx})\sigma_{zx}} \tag{4.1}$$

Equation 4.1 allows one to compute the bias of $R_{yx.z}$ with respect to the target quantity λ_{xy}, for any given hypothetical value of ε_{zx}, if the true model were given by G_A. Similarly, it allows one to determine how strong the unobserved confounder would need to be (as parameterized by ε_{zx}) such that the association $R_{yx.z}$ is completely explained by the unobserved confounder (i.e., the value of ε_{zx} such that $\lambda_{xy}(\varepsilon_{zx}) = 0$).

Still, what if the analyst has no knowledge to plausibly bound the strength of ε_{zx}? Even though the violation introduced in model G_A was the addition of the bidirected edge $x \leftrightarrow z$, corresponding to ε_{zx}, there is no reason to limit our attention to that parameter, and *any* θ-specific estimand could be used for sensitivity analysis. In fact, the two remaining parameters of the model also yield valid θ-specific estimands (Section 4.5 provides an algorithmic solution),

$$\lambda_{xy}(\lambda_{zx}) = \frac{\sigma_{xy} - \lambda_{zx}\sigma_{yz}}{1 - \lambda_{zx}\sigma_{zx}} \tag{4.2}$$

$$\lambda_{xy}(\varepsilon_{zy}) = \frac{\sigma_{zy} - \varepsilon_{zy}}{\sigma_{zx}} \tag{4.3}$$

Having a diverse option of sensitivity curves is important, because sensitivity analysis relies on plausibility judgments. One could argue, for instance, that assessing the plausibility of ε_{zx} could be hard because it involves judging the effect of confounders of *unknown* cardinality, and perhaps, previous studies give plausible bounds on the direct causal effect of z on x (i.e., λ_{zx}), making a λ_{zx}-specific estimand more attractive. Regardless of the specific scenario, it is clear that the choice of sensitivity parameters should be guided by the availability of substantive knowledge.

Remarkably, several subtleties arise when deriving θ-specific estimands, even in simple models with three variables. For instance, a natural approach for tackling the problem in our example could be the re-expression of $R_{yx.z}$ in terms of the covariance matrix implied by G_A, yielding,

$$R_{yx.z} = \lambda_{xy} - \frac{(\sigma_{zx} - \lambda_{zx})\varepsilon_{zy}}{1 - \sigma_{zx}^2} = \lambda_{xy} - \frac{\varepsilon_{zx}\varepsilon_{zy}}{1 - \sigma_{zx}^2} \tag{4.4}$$

One may surmise upon the examination of such expression that two sensitivity parameters are needed. As shown in Equations 4.1 to 4.3, this conclusion would be misleading.

These subtleties also appear when solving several variations of a model. Imagine the alternative model is now G_B, instead of G_A, as shown in Figure 4.1c. Is ε_{zx} an admissible sensitivity parameter in this case? Is the ε_{zy}-specific estimand derived in G_A still valid if the model were G_B? If we include another violation in both models, a bidirected arrow $x \leftrightarrow y$, would the previously obtained ε_{zy}-specific estimands still be valid? Despite the apparent similarity of both models, the answers to these questions reveal their sensitivity curves behave quite differently. The tools developed in this chapter not only provide an algorithmic solution to these questions, but also allow researchers to swiftly answer them by simple inspection of the graph.

The above examples demonstrate several of the identification problems entailed by a sensitivity analysis. If in small models these tasks are already complex, once we move to models with more than three or four variables, an informal, case-by-case approach to sensitivity analysis is simply infeasible. Therefore, we need a formal framework and efficient algorithms to incorporate constraints in linear SCMs.

4.4 Incorporating constraints in linear SCMs

Existing methods for identification in linear SCMs, such as the QID algorithm from [28], are able to incorporate constraints on directed edges and can be used to derive sensitivity curves such as the λ_{zx}-specific estimand of Equation 4.2. The QID algorithm exploits a known edge λ_{ab} by creating an auxiliary variable (AV) $b^* = b - \lambda_{ab}a$ [30]. Subtracting out the direct effect of a on b in this way may help with the identification of other coefficients in the model. For instance, the λ_{zx}-specific estimand can be computed using AVs in the following way: (i) create $x^* = x - \lambda_{zx}z$; (ii) use x^* as an instrument for λ_{xy}, resulting in $\lambda_{xy}(\lambda_{zx}) = \sigma_{yx^*}/\sigma_{xx^*} = (\sigma_{xy} - \lambda_{zx}\sigma_{yz})/(1 - \lambda_{zx}\sigma_{zx})$.[5]

However, neither the ε_{zx}-specific nor the ε_{zy}-specific estimands can be derived using QID; in fact, there is no current identification algorithm that offers a principled and efficient way to exploit knowledge of bidirected edges.[6] As this is critical for the derivation of sensitivity curves (see Section 4.6), one of the core contributions of this work is the development of a novel graphical procedure that allows one to systematically incorporate constraints on error covariances.

Conventional linear SCMs already impose one type of constraint on error covariances: a lack of a bidirected edge between two variables a and b encodes the assumption that the structural parameter ε_{ab} is zero. The identification problem imposed by sensitivity analysis, nonetheless, sets a different type of constraint—the error covariance ε_{ab} is fixed to a *known* but *non-zero* number. The essence of our method is to represent this knowledge in the graph.

Considering a graph G, covariance matrix Σ, and a *known* error covariance ε_{ab}, our strategy consists of performing a "manageable" transformation of G such that the bidirected edge $a \leftrightarrow b$ is removed from the graph. By "manageable" we mean the implied covariance matrix Σ'

[5] The QID algorithm extends generalized instrumental sets [20] using a bootstrapping procedure whereby complex models can be identified by iteratively identifying coefficients and using them to generate new auxiliary variables. It takes as inputs a graph G, covariance matrix Σ and *known* directed edges $\mathcal{D}$, and it returns the new set of identified directed edges.

[6] Methods from computer algebra offer a complete solution but are computationally intractable. See Section 4.6 and the discussion in the appendix.

of the transformed graph G' can still be derived from Σ and the known value ε_{ab}; otherwise, we would have no connection between G' and the data, making inference in G' impossible. Once this graphical transformation is applied, we can exploit *any* existing graphical identification method on the modified model G', and solutions in G' can be transfered back to solutions in the original model G. In short, we manipulate the graph to reduce an identification problem with a non-zero constraint to a standard one.

The easiest way to introduce our method, which we call PushForward, is via an example. Consider again graph G_A in Figure 4.1b, and assume ε_{zy} is known. Path-tracing [146] results in the following covariances, where the known parameter ε_{zy} is highlighted in red,

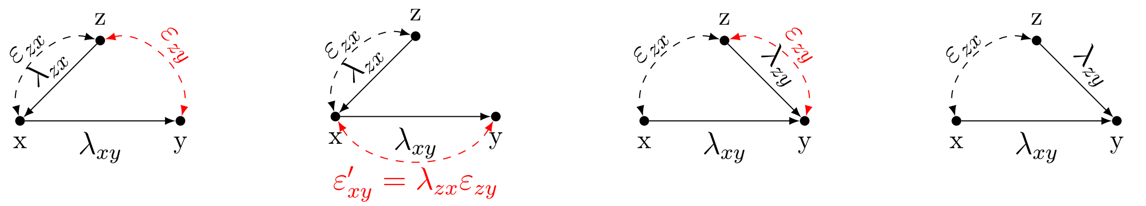
$$\sigma_{zx} = \lambda_{zx} + \varepsilon_{zx} \tag{4.5}$$

$$\sigma_{zy} = \lambda_{zx}\lambda_{xy} + \varepsilon_{zx}\lambda_{xy} + \varepsilon_{zy} \tag{4.6}$$

$$\sigma_{xy} = \lambda_{xy} + \lambda_{zx}\varepsilon_{zy} \tag{4.7}$$

Ideally, we could create an alternative model G^*_A where the bidirected edge $z \leftrightarrow y$ is fully removed from the graph. For this to be useful, we need to be able to express the new implied covariance matrix Σ^*_A in terms of the original covariance matrix Σ_A and the known error covariance ε_{zy}. While expressing σ^*_{zy} in terms of Σ_A and ε_{zy} is straightforward (since, trivially, $\sigma^*_{zy} = \sigma_{zy} - \varepsilon_{zy}$), it is not immediately clear how to write $\sigma^*_{xy} = \sigma_{xy} - \lambda_{zx}\varepsilon_{zy} = \lambda_{xy}$ in terms of Σ_A and ε_{zy}, for this requires identifying either λ_{xy} or λ_{zx} in the original model.

Thus, rather than fully removing $z \leftrightarrow y$, we "push it forward" to the children of z, as shown in graph G'_A of Figure 4.2b. Note the bidirected edge is moved from being between z and y to being between x (a child of z) and y, with new structural parameter $\varepsilon'_{zy} = \lambda_{zx}\varepsilon_{zy}$. Path-tracing of G'_A shows its implied covariance matrix Σ'_A is exactly the same as Σ_A, except



(a) G_A (b) G'_A (c) G_B (d) G'_B

Figure 4.2: Pushing forward ε_{zy} in G_A renders z a valid instrument in G'_A. Pushing forward ε_{zy} in G_B renders z single-door admissible in G'_B.

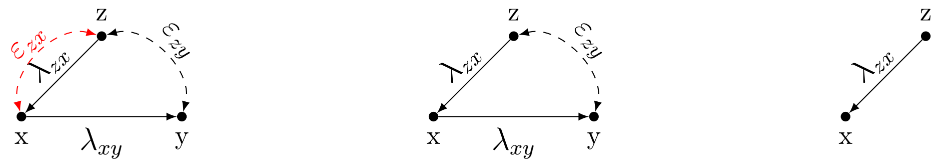
for σ'_{zy}, which can be obtained by subtracting ε_{zy} from σ_{zy},

$$\begin{aligned} \sigma'_{zx} &:= \sigma_{zx} && = \lambda_{zx} + \varepsilon_{zx} && (4.8) \\ \sigma'_{zy} &:= \sigma_{zy} - \varepsilon_{zy} && = \lambda_{zx}\lambda_{xy} + \lambda_{xy}\varepsilon_{zx} && (4.9) \\ \sigma'_{xy} &:= \sigma_{xy} && = \lambda_{xy} + \lambda_{zx}\varepsilon_{zy} && (4.10) \end{aligned}$$

Since G'_A has the same structural coefficients as G and we know how to compute the covariance matrix induced by G'_A from the known values Σ and ε_{zy}, we can use G'_A to identify the coefficients in our original model. In this case, z is an *instrument* for λ_{xy} in G'_A, resulting in the estimand $\lambda_{xy}(\varepsilon_{zy}) = \sigma'_{zy}/\sigma'_{zx} = (\sigma_{zy} - \varepsilon_{zy})/\sigma_{zx}$ of Equation 4.3.

Applying the same logic to graph G_B in Figure 4.1c, assume ε_{zy} is known. Since z has no other descendants except y, pushing forward ε_{zy} simply removes the bidirected edge $z \leftrightarrow y$. This results in the modified graph G'_B of Figure 4.2d with the amortized covariance of z and y, $\sigma'_{zy} = \sigma_{zy} - \varepsilon_{zy}$. Note ε_{zy} enters in no other covariances of the system. The graph G'_B renders z single-door admissible for the identification of λ_{xy}, giving us the estimand $\lambda_{xy}(\varepsilon_{zy}) = R'_{yx.z} = (\sigma_{yx} - \sigma_{xz}(\sigma_{zy} - \varepsilon_{zy}))/(1 - \sigma^2_{xz})$.

This simple graphical manipulation also makes it clear why adding a bidirected edge $x \leftrightarrow y$ as a further violation in the original graphs G_A and G_B has different consequences for the identification of λ_{xy}. In G'_A, z still remains a valid instrument even if the original graph had $x \leftrightarrow y$; this would only change the value of the structural coefficient ε'_{xy}, which would now read $\varepsilon'_{xy} = \varepsilon_{xy} + \lambda_{zx}\varepsilon_{zy}$. In G'_B, however, adding $x \leftrightarrow y$ renders z inadmissible for



(a) G_A (b) Push forward ε_{zx} (c) Prune y

Figure 4.3: Pushing forward ε_{zx} in G_A requires adjusting σ_{yz}. If the adjustment is possible, y is kept in the graph as in Figure 4.3b; if not, y is marginalized (pruned) as in Figure 4.3c.

single-door identification of λ_{xy}, since this backdoor path cannot be blocked.

Sometimes it might be necessary to prune variables from G' to guarantee Σ' is computable. Consider again G_A and assume ε_{zx} is known. Pushing forward ε_{zx} results in Figure 4.3b where, as before, we know $\sigma'_{zx} = \sigma_{zx} - \varepsilon_{zx}$. However, path-tracing of Figure 4.3b shows the covariance of z with y would also need adjustment, $\sigma'_{zy} = \sigma_{zy} - \lambda_{xy}\varepsilon_{zx}$. Thus, we have two cases: (i) if λ_{xy} is known, the adjustment is feasible and we are done; (ii) if λ_{xy} is not (yet) known, the adjustment cannot be made; but, since y is a leaf node, it can be pruned from G' [138], avoiding this problem (Figure 4.3c). In this case, note the pruned graph is still helpful—now λ_{zx} can be identified. As previously discussed, knowledge of λ_{zx} permits identification of λ_{xy} using AVs, giving us the ε_{zx}-specific estimand of Equation 4.1.

The graphical manipulation of PushForward is general, and can be performed whenever we have knowledge of a known error covariance. Theorem 1 formalizes the procedure to arbitrary models. Given any bidirected edge $x \leftrightarrow y$ with known value ε_{xy}, we remove it from the graph and register the new amortized covariance $\sigma'_{xy} = \sigma_{xy} - \varepsilon_{xy}$. Next we repair the covariances of the descendants of x with y by, for every $c \in Ch(x)$, adding (or modifying) the bidirected edge $c \leftrightarrow y$ with the direct causal effect λ_{xc} times ε_{xy}. Finally, for any descendant z of y, we either (i) amortize its covariance with x, if *all* edges that compose the total causal effect δ_{yz} of y on its descendant z are known, or (ii) marginalize z out by pruning the graph. The final output is a modified model $\langle G', \Sigma' \rangle$ where any graphical identification method can be applied; and, estimands in terms of Σ' can be converted back to estimands in terms of Σ and ε_{xy}.

Theorem 1 (PushForward)**.** *Given a linear SCM with graph G, covariance matrix Σ, a set* *of known directed edges $\mathcal{D}$, and known bidirected edge ε_{xy}, let the pair $\langle G', \Sigma' \rangle$ be constructed from G and Σ as follows:*

1. *$x \leftrightarrow y$ is removed and $\sigma'_{xy} = \sigma_{xy} - \varepsilon_{xy}$;*
2. *$\forall c \in Ch(x), c \neq y$, the bidirected edges $c \leftrightarrow y$ are added if they do not exist, and $\varepsilon'_{cy} = \varepsilon_{cy} + \lambda_{xc}\varepsilon_{xy}$;*
3. *$\forall z \in De(y)$, $z \neq x$, if there is an edge on any directed path from y to z that is not in $\mathcal{D}$, then z is removed from G'. For the remaining z, $\sigma'_{xz} = \sigma_{xz} - \varepsilon_{xy}\delta_{yz}$, where δ_{yz} is the sum of all directed paths from y to z;*
4. *All other parameters and covariances remain the same.*

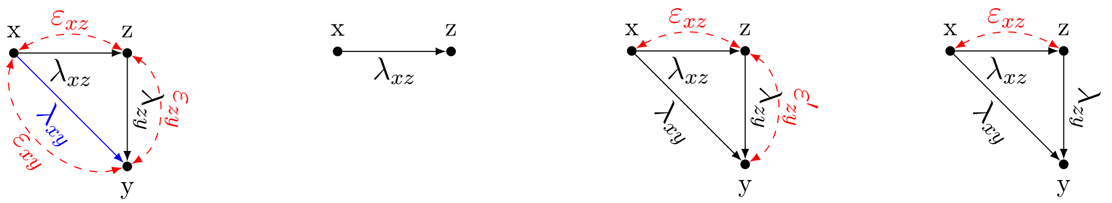
Then, if λ_{ab} is identifiable in G', it is $(\varepsilon_{xy}, \mathcal{D})$-identifiable in G.

We denote by PF$(G, \Sigma, \mathcal{D}, \varepsilon_{xy}, x)$ the function that returns the modified model $\langle G', \Sigma' \rangle$ as per Theorem 1. Pseudocode for PF (which closely follows the steps of the theorem) as well as the proof can be found in the appendix.

4.5 Algorithmic derivation of sensitivity curves

In this section, we construct a graph-based constrained identification algorithm for linear SCMs which systematically exploits knowledge of *both* path coefficients and error covariances efficiently. Our algorithm relies on the PushForward method to incorporate constraints on bidirected edges, and on the AV technique (via the qID algorithm) to incorporate constraints on directed edges. This allows the algorithmic derivation of sensitivity curves for a target query λ_{xy} in arbitrary linear models, with an arbitrary set of directed and bidirected edges as sensitivity parameters.

Although the graphical modification of PushForward is defined for one bidirected edge, the modified graph G' is a valid model in which any graphical operation can be performed. We can thus extend PushForward to handle multiple bidirected edges by iteratively applying



(a) G (b) $G_z^{'}$ (c) $G_y^{'}$ (d) $G_y^{''}$

Figure 4.4: PF multiple edges in topological order.

it whenever a bidirected edge of the modified graph is still known—what remains to be decided is the order in which to perform these operations. Note that testing all possible orders of graphical manipulations can result in an algorithm with exponential computational complexity, even when initially pushing forward a single bidirected edge ε_{xy}. This happens because new bidirected edges are created for each $c \in Ch(x)$ and, if all the λ_{xc} are identifiable, all subsets of those bidirected edges may be eligible to be pushed forward again. Thus, here we propose an efficient procedure using topological ordering, which performed *as well* as a brute-force approach in our computational experiments (Section 4.6).

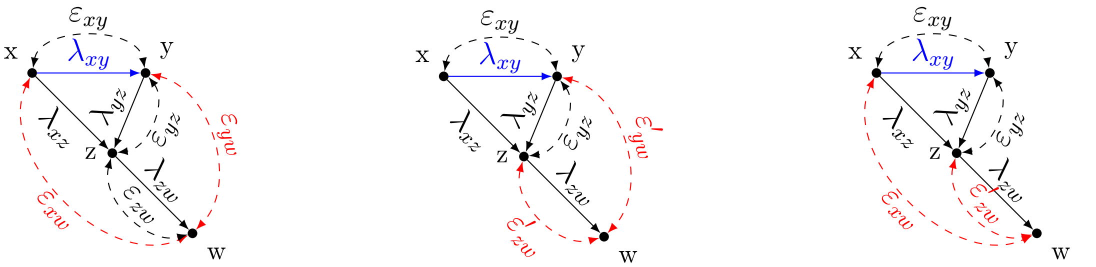
Consider the example given in Figure 4.4a. The task is to decide whether $\theta = (\varepsilon_{xz}, \varepsilon_{xy}, \varepsilon_{zy})$ (in red) is an admissible set of sensitivity parameters for the target coefficient λ_{xy} (in blue) and, if so, to find the corresponding sensitivity curve. Our strategy consists of, for each node v, listing its ancestors $a \in An(v)$, and, in topological order, iteratively push forward ε_{av} if it is still known in the modified graph. By performing operations in this way, we are guaranteed to visit each ancestor of v only once. Starting with node $v = z$, it has only one ancestor x and a single known bidirected edge to be removed, ε_{xz}. This can be handled with a one-step PushForward operation (pruning y), resulting in the modified graph G_z' of Figure 4.4b, in which λ_{xz} can be trivially identified. Next, return to the original graph and consider $v = y$, with ancestors x and z. Following a topological order, we first push forward ε_{xy}, giving us the modified graph $G_y^{'}$ of Figure 4.4c with new bidirected edge $\varepsilon'_{zy} = \varepsilon_{zy} + \lambda_{xz}\varepsilon_{xy}$. Note all components of ε'_{zy} are known, we can thus push forward ε'_{zy} in G_y', obtaining the graph $G_y^{''}$ in Figure 4.4d, in which λ_{xy} is identified with sensitivity curve $R''_{yx.z}$.

In the previous example we demonstrated how to systematically deal with bidirected

Algorithm 1 $\text{cID}(G, \Sigma, \mathcal{D}, \mathcal{B})$

```
initialize $V_{\mathcal{B}} \leftarrow \text{Vertices}(\mathcal{B})$
repeat
  $\mathcal{D} \leftarrow \mathcal{D} \cup \text{qID}(G, \Sigma, \mathcal{D})$
  for each $v \in V_{\mathcal{B}}$ do
    $\langle G', \Sigma' \rangle \leftarrow \langle G, \Sigma \rangle$
    for each $a \in An(v)$ in topological order do
      if $\varepsilon'_{av}$ is known then
        $\langle G', \Sigma' \rangle \leftarrow \text{PF}(G', \Sigma', \mathcal{D}, \varepsilon'_{av}, a)$
        $\mathcal{D} \leftarrow \mathcal{D} \cup \text{qID}(G', \Sigma', \mathcal{D})$
      end if
    end for
  end for
  for each $\varepsilon_{ab} \in \mathcal{B}$ do
    $\langle G', \Sigma' \rangle \leftarrow \text{PF}(G, \Sigma, \mathcal{D}, \varepsilon_{ab}, a)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \text{qID}(G', \Sigma', \mathcal{D})$
  end for
until all directed edges have been identified or no edge has been identified in the last
iteration
```

edges connected to *ancestors* of a node v; however, in linear models, *descendants* of v can also help with the identification of direct causal effects λ_{av}. Consider, for instance, Figure 4.5a. The task is to find a sensitivity curve for λ_{xy} in terms of $\theta = (\varepsilon_{xw}, \varepsilon_{yw})$. Start with node w and, as before, push forward ε_{xw} as in Figure 4.5b. Here, λ_{zw} can be identified with x as an instrument. Returning to the original graph, now consider node y and push forward the bidirected edge ε_{yw} with its *descendant* w, as in Figure 4.5c. Since λ_{zw} has been identified, we can create the AV $w^* = w - \lambda_{zw} z$ which is a valid instrument for λ_{xy}.

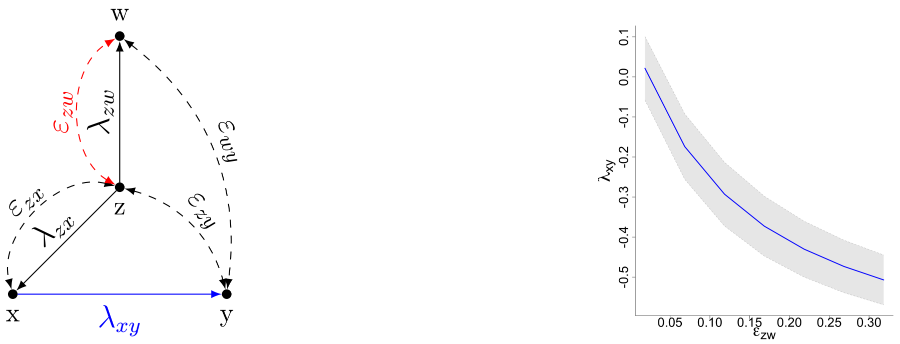


(a) Original Model

(b) Push forward ε_{xw}

(c) Push forward ε_{yw}

Figure 4.5: Instruments with ancestors and descendants.



(a) λ_{xy} is ε_{zw}-identifiable

(b) Sensitivity of λ_{xy} in terms of ε_{zw}

Figure 4.6: In Fig 4.6a note that, although not connected to x nor y, ε_{zw} is an admissible sensitivity parameter for λ_{xy}. Figure 4.6b shows the sensitivity curve of λ_{xy} in terms of ε_{zw} for a numerical simulation of the model in Figure 4.6a.

These two cases illustrate our general procedure for handling multiple bidirected edges, which in combination with the QID algorithm forms our constrained identification algorithm cID, provided in Algorithm 1. Lines 4 to 12 perform PushForward (PF) in topological ordering, each time applying QID in the modified model to verify if new directed edges can be identified; lines 13 to 16 perform a single PushForward operation on each bidirected edge, which may free descendants to be used as instruments as in Figure 4.5. Since new identified edges can help both PushForward as well as QID, this process is repeated until all or no new directed edges are identified in the last iteration. The complexity of cID is dominated by QID, which is polynomial if the degree of each node is bounded [28].

An interesting 4-node example is shown in Figure 4.6a, where ε_{zw}, a parameter neither related to x nor y, is an admissible sensitivity parameter for λ_{xy}! Our algorithm derives an ε_{zw}-specific estimand for λ_{xy} as follows. It first pushes forward ε_{zw}, and runs QID in the modified graph, resulting in the identification of λ_{zw}. Next, the algorithm returns to the original graph, and runs QID, which uses λ_{zw} to create the auxiliary variable $w^* = w - \lambda_{zw} z$, enabling the identification of λ_{zx}. Finally, still within QID, λ_{xy} is obtained using the auxiliary variable $x^* = x - \lambda_{zx} z$.

As discussed in Section4.3, the utility of θ-specific estimands is to show how sensitive the target quantity of interest is to different hypothetical values of the sensitivity parameters θ.

These results can then be submitted to quantitative plausibility judgments, for instance, in the form of $\theta \in \Theta_p$, where Θ_p is a plausibility region. To illustrate how one could deploy this in practice, we provide a numerical example of the causal model in Figure 4.6a. Our goal is to assess how different hypothetical values for ε_{zw} affects inference of λ_{xy}. In a real context, this needs to be estimated from finite samples, and here we use a maximum likelihood estimator. Figure 4.6b shows the estimates for λ_{xy} (blue) for different values of the sensitivity parameter ε_{zw}, along with the corresponding 95% confidence interval (gray). If, for instance, we can plausibly bound ε_{zw} to be within 0.1 to 0.3, the plot reveals λ_{xy} can be safely judged to be within -0.2 to -0.6.

4.6 Computational experiments

The identification problem in linear systems has not yet been efficiently solved. Although there exists a complete solution using computer algebra [66], these methods are computationally intractable, making it impractical for graphs larger than 4 or 5 nodes. Since we rely on existing identification algorithms that are polynomial but not complete (i.e., QID cannot find all identifiable parameters), we cannot expect the CID algorithm to find all sensitivity curves as well. In this section, we report the results of an extensive set of experiments aimed to empirically verify the generality of our approach. We have performed an exhaustive study of all possible queries in 3 and 4-node models, which are essentially the largest instances computer algebra methods can solve through brute force.[7]

A query consists of determining whether in model G, a target parameter λ_{xy} is θ-identifiable given a set of sensitivity parameters θ. For 3-node models, we have 50 connected graphs with 720 possible queries; for 4-node models, we have 3,745 connected graphs and 1,059,156 possible queries.[8] For each query, we used algebraic methods to determine ground-truth

[7] We use Gröbner bases, which has a doubly-exponential computational complexity [14]. See appendix for details.

[8] For 5-node models, these numbers reach 1 million graphs and 11 *billion* queries. Ground-truth computations in 5-node models using computer algebra can take hours for a *single* graph.

	3-Node Models			4-Node Models		
ID Algorithm	*Directed*	*Bidirected*	*Both*	*Directed*	*Bidirected*	*Both*
qID (AVs only)	19(100%)	– (0%)	68 (21%)	14,952(95%)	– (0%)	170,304(29%)
cID (AVs + PF)	19(100%)	109(100%)	320(100%)	14,952(95%)	50,708(97%)	555,758(96%)
Ground Truth	19	109	320	15,740	52,016	578,858

Table 4.1: Number of θ-identifiable sensitivity queries (only when $\theta \neq \emptyset$) per type of sensitivity parameters θ.

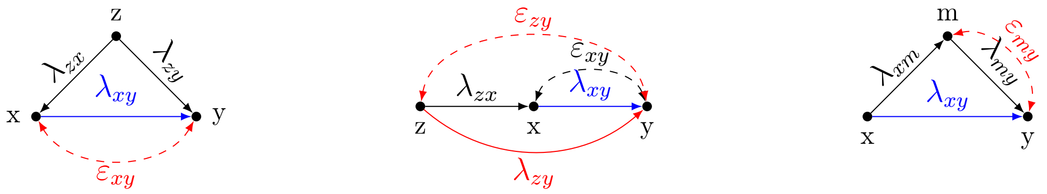
identification and checked it against the results of both qID and cID. Our interest lies in the queries that are θ-identifiable *only* when $\theta \neq \emptyset$.

The results are given in Table 4.1, where columns restrict sensitivity parameters θ to be: (i) subsets of directed edges; (ii) subsets of bidirected edges; and, (iii) subsets of both directed and bidirected edges. The results show that our cID algorithm correctly identifies all possible sensitivity curves for 3-node models. Among 4-node models, our method solves 96% of all identifiable sensitivity queries.

These numbers reveal that, in the context of *linear* SCMs, canonical sensitivity analysis examples which have been addressed on a case-by-case basis in the literature (e.g., Figure 4.7, target coefficient in blue and sensitivity parameters in red), are only a *small* subset of all possible sensitivity analyses exercises enabled by our proposal. When comparing cID's results to those of qID only, it is also clear that systematically incorporating constraints on bidirected edges is essential for obtaining sensitivity curves.

A valid concern regarding cID's current implementation is that the proposed topological ordering for processing bidirected edges could be less capable than a general search over all possible valid graphical manipulations. With this in mind, we performed a thorough comparison of our proposal against other ordering methods for all queries in 3 and 4-node models. Topological ordering proved to perform *as well* as a brute-force search that recursively tests all possible subsets of bidirected edges that can be pushed forward.

Finally, the incompleteness of cID can stem from two sources: limitations of the graphical manipulations performed by PushForward and the incompleteness of the identification



(a) Backdoor

(b) IV

(c) Mediation

Figure 4.7: Canonical sensitivity analyses: (a) backdoor violation with unobserved confounders independent of observed confounders [25]; (b) putative instrumental variable, where both the exclusion and independence restriction are suspected to be violated [143]; (c) randomized trial in which treatment x has side-effect m, and unobserved mediation-outcome confounding cannot be ruled out [140]. For linear SCMs, these are special cases of all queries solved by our approach.

algorithm for directed edges, QID. Separating the two can help guide efforts for future research. To achieve that, we used algebraic methods to simulate how CID would have performed if it had access to a complete identification algorithm for directed edges instead of QID. We found that CID would have identified over 99.99% of 4-node sensitivity queries. This seem to suggest that: (i) the main bottleneck of CID is QID; and (ii) PUSHFORWARD with topological ordering can reap the benefits of improved identification algorithms for directed edges.

4.7 Conclusion

In this chapter, we introduced a general algorithmic framework of sensitivity analysis for linear SCMs. We reduced sensitivity analysis to a constrained identification problem and developed a novel graphical procedure to systematically incorporate constraints on bidirected edges. We then devised an efficient graph-based algorithm for deriving sensitivity curves. Exhaustive experiments corroborated the generality of our proposal. Such systematic tools can help analysts better navigate in the model space and understand the trade-off between the plausibility of assumptions and the strength of conclusions. Extensions to other types of violations and to nonlinear models are promising directions for future work.

CHAPTER 5

Generalizing Experimental Results by Leveraging Knowledge of Mechanisms

5.1 Introduction

Generalizing results of randomized control trials (RCT) is critical in many empirical sciences and demands an understanding of the conditions under which such generalizations are feasible. When the mechanisms that determine the outcome differ between the study population and the target population, generalization requires measuring the variables responsible for such differences or, if this is not possible, isolating them away by measuring other variables [113]. Recent work [78, 79, 77] describes an interesting situation under which transportability across populations is feasible without such measurements. This feasibility, however, is not immediately inferable using a standard (non-parametric) selection diagram [113, 16], because it relies on the invariance of only some components of the outcome mechanism, but not all.

In this chapter, we use the theory of Structural Causal Models (SCM) [109] to show how generalization in these settings can be modeled using ordinary structural equations, counterfactual logic and selection diagrams. We demonstrate that it requires two key assumptions: (i) the independence of causal factors that affect the outcome; and, (ii) *functional constraints* on how these factors interact to produce the outcome. The combination of these assumptions may entail the invariance of certain *probabilities of causation* [108, 137] across domains, thus allowing the transport of causal effects in settings where non-parametric generalization is otherwise impossible.

We further extend the results of existing literature by: (i) relaxing the monotonicity assumption and providing bounds for the causal effect in the target domain; (ii) deriving novel identification and over-identification results for probabilities of causation, as well as the transported causal effect, when trials from multiple source domains are available; and, (iii) providing a Bayesian framework for estimating the transported causal effect from finite samples. We illustrate these methods both in simulated data and in a real example that generalizes the effects of Vitamin A supplementation on childhood mortality across different regions [132, 102, 144]. Open source software for `R` implements the methods discussed in this chapter.[1]

5.2 Motivating example

To fix ideas, we borrow the "Russian Roulette" example from [77]. Although stylized, this intuitive example illustrates the key features of the problem.

5.2.1 A Russian Roulette trial

Suppose the city of Los Angeles decides to run a randomized control trial (RCT) to assess the effect of playing "Russian Roulette" on mortality.[2] After running the experiment, the mayor of Los Angeles discovers that "Russian Roulette" is harmful: among those assigned to play Russian Roulette, 17.5% of the people died, as compared to only 1% among those who were not assigned to play the game (people can die due to other causes during the trial, for example, prior poor health conditions).

After hearing the news about the Los Angeles experiment, the mayor of New York City (a dictator) wonders what the overall mortality rate would be if the city forced everyone to play Russian Roulette. Currently, the practice of Russian Roulette is forbidden in New York, and its mortality rate is at 5% (4% higher than LA). The mayor thus asks the city's statistician

[1] Available in `https://github.com/carloscinelli/generalizing`.

[2] Russian Roulette consists of loading a bullet into a revolver, spinning the cylinder, pointing the gun at one's own head and then pulling the trigger. We do not recommend attempting this.

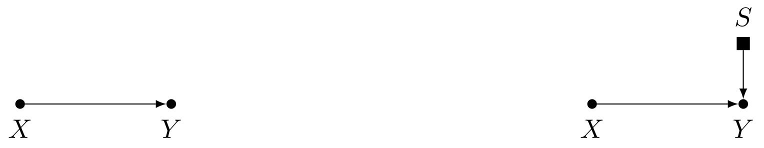
to decide *whether* and *how* one could use the data from from Los Angeles to predict the mortality rate in New York, once the new policy is implemented.

Intuitively, our causal knowledge of the domain permits us to answer the question posed by the NYC mayor. Mortality is a consequence of two "independent" processes (the game of Russian Roulette and prior health conditions of the individual), and while the first factor remains unaltered across cities, the second intensifies by a known amount (5% vs 1%). Moreover, we can safely assume that the two processes interact disjunctively, namely, that death occurs if and only if at least one of the two processes takes effect. From these two assumptions and elementary probability theory, we can conclude that mortality in NYC would be 20.8%. In section 5.3 we will cast this intuition into a formal setting, define this notion of "independence," and show how the data from NYC and LA should be combined to match our expectation. But before that, let us examine how this intuition clashes with the conclusion of a coarse analysis using selection diagrams.

5.2.2 An "impossibility" result

Selection diagrams are causal diagrams enriched with "selection nodes" S, usually represented by square nodes (■). These new nodes are used by the analyst to indicate which *local mechanisms* are suspected to differ between two environments (in our example, the mortality mechanism is suspected to differ between Los Angeles and New York). More importantly, the absence of a selection node pointing to a variable represents the *assumption* that the local mechanism responsible for assigning the value to that variable is the same in the two populations [106, 109, 113, 16].

To build our selection diagram, we need to introduce some notation. The population of Los Angeles will be denoted by Π (the "source population") and that of New York by Π^* (the "target population"). The random variable Y stands for mortality, with events $Y = 1$ denoting "death" and $Y = 0$ denoting "survival;" the random variable X stands for the "treatment" assignment, with events $X = 1$ denoting "play Russian Roulette" and $X = 0$ denoting "not play Russian Roulette." The random variable Y_x denotes the potential response of Y when the treatment X is experimentally set to x. Thus, mathematically, the findings of the RCT can be translated to $P(Y_1 = 1) = 17.5\%$ and $P(Y_0 = 1) = 1\%$, and the available data from New York is $P^*(Y_0 = 1) = 5\%$. Our task is to estimate $P^*(Y_1 = 1)$.



(a) Coarse causal diagram

(b) Coarse selection diagram

Figure 5.1: Coarse causal (a) and selection (b) diagrams of the Russian Roulette trial. The presence of $S \to Y$ in (b) correctly prohibits the naive transportation of the interventional distribution $P(Y_x)$ from the source Π (Los Angeles) to the target environment Π^* (New York).

The coarsest causal diagram of the Russian Roulette trial comprises only the treatment X and the outcome Y, as shown in Figure 5.1a. To move from the causal diagram to the selection diagram, we need to think of what may differ between LA and NYC. Since we already know from the data that $P(Y_0 = 1) \neq P^*(Y_0 = 1)$, we suspect there are differences in the way mortality is determined in the two cities (for example, people in New York may be in poorer health conditions, or the air quality may be worse). Thus, the selection diagram must contain a selection node S pointing to the mortality variable Y to indicate this disparity, as shown in Figure 5.1b.

Graphically, checking whether a causal relationship is transportable from one environment to another involves checking whether there exists a set of measurements that d-separates [109] the source of disparity (the selection node S) from our target quantity. The presence of the selection node pointing directly into Y prevents the separation of S from Y, and leads us to conclude that transportability is impossible without further assumptions. On the other hand, the intuition that led us to predict the new mortality rate in NYC tells us that such assumptions, once formalized, could license transportability. This intuition, as we discussed, was based on two assumptions that are not shown in the coarse selection diagram of Figure 5.1. The diagram represents only the existence of a disparity between LA and NYC, not the fact that it is localized to one cause of death (prior health factors), and that it does not extend to the other cause (the game of Russian Roulette). As a result, the diagram correctly warns us that, absent further assumptions, we are not authorized to make any generalization between the two cities.

5.3 Building the structural model

We now explicate formally what we know about the game of "Russian Roulette" and health factors, and show how this knowledge renders transportability possible.

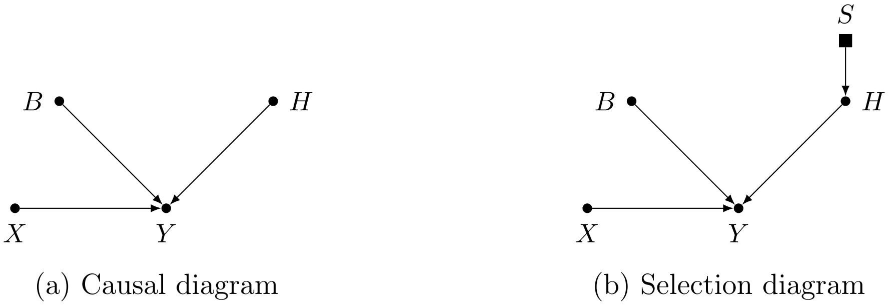
5.3.1 Prior health conditions *versus* physical mechanism

To represent the two causes of death, we refine our model by defining two extra random variables, B and H: (i) B denotes "bad luck" when playing Russian Roulette, and its values represent a match ($B = 1$) or mismatch ($B = 0$) between the trigger and the location of the bullet in the cylinder; (ii) and H denotes *all* other health factors producing death ($H = 1$) or survival ($H = 0$). Accordingly, our causal diagram will contain two new edges, $H \rightarrow Y$ and $B \rightarrow Y$, since both "health conditions" and "bad luck" are key determinants of mortality Y. The updated causal diagram is shown Figure 5.2a. Note the absence of a directed or bidirected edge between H and B, which encodes our assumption that these two mechanisms are activated independently of each other.[3]

The new model helps us see more clearly the commonalities and disparities between LA and NYC. First, since there is a multitude of factors that can affect prior health conditions, and those are likely to differ between the two cities (as suggested by the observed difference $P(Y_0 = 1) \neq P^*(Y_0 = 1)$), we again introduce a selection node pointing to H. Moreover, to encode the assumption that the probability of "bad luck" occurring is the same in both cities, we do not connect B to a selection node.[4] The new selection diagram is shown in Figure 2b.

[3] The arrow $X \rightarrow Y$ comprises, of course, many intermediate mechanisms (such as loading the gun, spinning the cylinder, pulling the trigger) that are not modeled explicitly.

[4] Note that, although reasonable, one cannot take this assumption for granted—it could be the case that revolvers used for Russian Roulette in New York have a different number of chambers than those used in Los Angeles. *The absence of a selection node* pointing to B encodes the *assumption* that this is not the case.



(a) Causal diagram (b) Selection diagram

Figure 5.2: New causal (a) and selection (b) diagrams explicitly including the variables "health conditions" (H) and "bad luck" (B) when playing Russian Roulette. Here the analyst asserts (using the selection node S) that H may differ between LA and NYC, but assumes that the mechanism triggering B is the same between the two cities. Also important is the absence of a directed edge or a bidirected edge between H and B.

The diagram of Figure 5.2b now guides us toward leveraging the data obtained in LA to make predictions in NYC. If we can find a way to *block the source of disparity originating from* H, we would be left with the invariant physical mechanism shared by both cities. However, since H is unobserved, blockage is impossible without further assumptions. We now ask whether our understanding of how the two mechanisms interact in producing Y would permit us to estimate $P^*(Y_1 = 1)$.

5.3.2 Leveraging functional constraints

Our understanding that mortality is caused by *either one* of the two processes (prior health conditions or bad luck in the game), dictates the following *functional specification* for the *structural equation* of Y,

$$Y = H \vee (X \wedge B) \tag{5.1}$$

Where $\vee$ denotes the logical "or" operator, and $\wedge$ denotes the logical "and" operator. Like any structural equation, Equation 5.1 defines the potential outcomes Y_0 and Y_1 [109, Ch.7] which we may now find useful to encode explicitly. Its first implication is that $Y_0 = H$ and $Y_1 = H \vee B = Y_0 \vee B$. This tells us that, once we know the potential response of units under no treatment (Y_0) we do not need to know anything else about their previous health

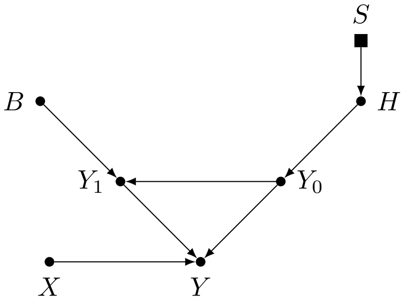


Figure 5.3: Selection diagram explicitly showing the potential outcomes Y_0 and Y_1 as implied by the functional constraints. Note that $Y_1 \perp\!\!\!\perp S \mid Y_0$.

condition (H) to determine the value of Y_1—B would suffice.[5] We can represent this fact in a modified selection diagram, in which the potential outcomes are now also shown explicitly (Figure 5.3). The diagram reveals that Y_0 blocks the source of health disparities between the two populations, and we conclude that $Y_1 \perp\!\!\!\perp S \mid Y_0$.[6]

More concretely, consider the counterfactual quantity

$$\mathrm{PS}_{01} := P(Y_1 = 1 \mid Y_0 = 0)$$

which stands for the share of people who would die if forced to play Russian Roulette, among those who would not have died if not forced to do so. In other words, PS_{01} represents the probability that the game of Russian Roulette is *sufficient* to *kill* a person *during the trial.* The acronym PS_{01} was chosen to emphasize its relation to the "probability of sufficiency" (PS), $\mathrm{PS} = P(Y_1 = 1|Y = 0, X = 0)$, as defined and analyzed in [108] and [137]. In our

[5] Although here we have $Y_0 = H$ for simplicity, this need not be the case. The same argument would hold, for instance, if we define H to be a random variable with arbitrary cardinality and $Y = g(H) \vee (X \wedge B)$, where $g(H) \in \{0, 1\}$. Likewise, see Appendix 7.4.1 for an example where the treatment variable X is continuous and the same strategy adopted here can be employed.

[6] Since some relationships in the graph may be deterministic, conditional independencies other than those revealed by d-separation (with lower-case d) may be present. A complete criterion for DAGs with deterministic nodes is given by the D-separation criterion (with capital D) of [67]. Moreover, note arrows between potential outcomes need not convey causal influence; their purpose is merely to ensure that the correct conditional independencies among variables are encoded in the graph, *as derived from the structural equations.* Finally, here we are not treating the question of how scientists acquire scientific knowledge in the form of a functional specification such as Equation 1. Rather, our task is more modest: given that scientists sometimes have knowledge of mechanisms, how can we leverage some of that knowledge for identification.

context, since the treatment is randomized, the two quantities coincide,

$$P(Y_1 = 1|Y_0 = 0) = P(Y_1 = 1|Y_0 = 0, X = 0) = P(Y_1 = 1|Y = 0, X = 0)$$

where the first equality is licensed by the randomization of X and the second equality is due to consistency. In general, however, PS_{01} need not be the same as PS—the later measures the probability of fatal treatment among those who, given the choice, would *choose* not to be treated and survive; the former measures the probability of fatal treatment among those who would survive had they not been *assigned* for treatment.[7] Similar reasoning holds for $\mathrm{PS}_{10} := P(Y_1 = 0 \mid Y_0 = 1)$, which stands for the probability that playing Russian Roulette is *sufficient* to *save* a person who would die if denied treatment. In our example, this probability is obviously zero as we shall formally show below. The condition $Y_1 \perp\!\!\!\perp S \mid Y_0$, implied by the diagram, states that these *probabilities of causation* are invariant across cities.[8] This feature of invariance, which is important in its own right, follows solely from our structural assumption about the mechanisms involved.

A second implication of Equation 5.1 is that the treatment effect is *monotonic*, that is $Y_1 \geq Y_0$ for all individuals. This, in turn, implies $\mathrm{PS}_{10} = 0$; in other words, an individual that would have died of other causes during the trial, would still die if forced to play Russian Roulette. It has been shown that monotonicity is sufficient for identifying PS_{01} in this setting [108, 137, 78]. Indeed, by the law of total probability,

$$P(Y_1 = 1) = (1 - \mathrm{PS}_{10})P(Y_0 = 1) + \mathrm{PS}_{01}(1 - P(Y_0 = 1))$$

The quantity $P(Y_0 = 1)$ is given from the RCT (1%) and, due to monotonicity, $\mathrm{PS}_{10} = 0$.

[7] For example, in legal settings, where acts are executed by *choice*, conditioning on the *observed* X gives a more appropriate measure of an agent's responsibility, as argued in [109] and [111].

[8] Probabilities of causation have been extensively studied elsewhere under a different context. See [108, 137, 109].

Thus, we have:

$$PS_{01} = \frac{P(Y_1 = 1) - P(Y_0 = 1)}{1 - P(Y_0 = 1)} = \frac{17.5\% - 1\%}{99\%} = 1/6$$

This is not surprising; the probability that the "treatment" is *sufficient* to kill an individual who would have otherwise survived indeed equals 1/6—the probability of having "bad luck" in the game of Russian Roulette, using a revolver with six chambers.[9]

Thus far we have established that $\mathrm{PS}_{10} = \mathrm{PS}^*_{10}$, $\mathrm{PS}_{01} = \mathrm{PS}^*_{01}$, and that $\mathrm{PS}_{10} = 0$, $\mathrm{PS}_{01} = 1/6$. Combining these results with the current baseline mortality from NYC, that is, $P^*(Y_0 = 1) = 5\%$, we can finally evaluate our target quantity $P^*(Y_1 = 1)$,

$$\begin{aligned}
P^*(Y_1 = 1) &= (1 - \mathrm{PS}^*_{10})P^*(Y_0 = 1) + \mathrm{PS}^*_{01}(1 - P^*(Y_0 = 1)) \\
&= (1 - \mathrm{PS}_{10})(5\%) + \mathrm{PS}_{01}(95\%) \\
&= (1)(5\%) + (1/6)(95\%) = 20.8\%
\end{aligned}$$

Which matches the intuitive answer obtained in Section 5.2.

As a brief remark, note that, if instead of $Y_1 \perp\!\!\!\perp S \mid Y_0$ we had obtained the condition $Y_0 \perp\!\!\!\perp S \mid Y_1$, we would conclude that the probabilities $\mathrm{PN}_{01} := P(Y_0 = 0 \mid Y_1 = 1)$ and $\mathrm{PN}_{10} := P(Y_0 = 0 \mid Y_1 = 1)$ are the same across trials. These quantities represent the probability that the treatment is *necessary* for causing (PN_{01}) or preventing (PN_{10}) the outcome during the experiment. All results of this chapter hold in this setting, with minor modifications. Therefore, for simplicity of exposition, in the remainder of the text we discuss the case of $Y_1 \perp\!\!\!\perp S \mid Y_0$ only.[10]

[9] The right-hand side of this expression is known as the "relative difference," or "susceptibility." Simple algebra shows that $\frac{P(Y_1=1)-P(Y_0=1)}{1-P(Y_0=1)} = 1 - \frac{1-P(Y_1=1)}{1-P(Y_0=1)}$, where the quantity $\frac{1-P(Y_1=1)}{1-P(Y_0=1)}$ is known as the "survival ratio." Since under the assumption of monotonicity these estimands identify PS_{01}, and PS_{01} is invariant across domains, it thus follows that the "relative difference" and the "survival ratio" will also be equal between populations. [78] suggested using this fact as a rationale for assuming homogeneity of effect measures across domains, a common heuristic among epidemiologists for approaching generalizability problems. These equivalences, however, break down without monotonicity; in that case, the "relative difference" is a lower bound for the probability of sufficiency [137], as we discuss next.

[10] For example, under the assumption of monotonicity, we have that $\mathrm{PN}_{01} = \frac{P(Y_1=1)-P(Y_0=1)}{P(Y_1=1)}$ [108].

5.3.3 Bounds without monotonicity

A key step in obtaining a point estimate for $P^*(Y_1 = 1)$ was the monotonicity property, which emanates from the functional form of Equation 5.1. Monotonicity allowed us to identify the probabilities of sufficiency PS_{01} and PS_{10}, which, as advertised by the assumptions in the selection diagram of Figure 5.3, are invariant across domains. The monotonicity property holds trivially in our example of the Russian Roulette, when Y represents death, but it may not hold for other outcomes or, more generally, it may not hold in contexts beyond our stylized example.

Remarkably, however, even in the absence of monotonicity, one can still assess the transported causal effect, albeit in the form of a *bound*. The next theorem shows that the counterfactual independence $Y_1 \perp\!\!\!\perp S \mid Y_0$ by itself is strong enough for bounding the causal effect in the target domain. These results improve the bias analysis performed by [78], and provide an exact characterization of the inferences compatible with the assumption of $Y_1 \perp\!\!\!\perp S \mid Y_0$.

Theorem 2. *Consider a source domain Π and a target domain Π^*. Let $P_{ij} := P(Y_i = j)$, $P^*_{ij} := P^*(Y_i = j)$, and let $RR = \frac{P_{11}}{P_{01}}$ denote the* risk-ratio *in the trial of the source domain Π. If $Y_1 \perp\!\!\!\perp S \mid Y_0$, then P^*_{11} of Π^* is bounded by $P^{*L}_{11} \leq P^*_{11} \leq P^{*U}_{11}$, with,*

$$P^{*L}_{11} = RR \times P^*_{01} + \min\left\{ \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right) PS^L_{01},\ \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right) PS^U_{01} \right\},$$

$$P^{*U}_{11} = RR \times P^*_{01} + \max\left\{ \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right) PS^L_{01},\ \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right) PS^U_{01} \right\}$$

where $PS^L_{01} = \max\left\{0, \frac{P_{11} - P_{01}}{1 - P_{01}}\right\}$ and $PS^U_{01} = \min\left\{\frac{P_{11}}{1 - P_{01}}, 1\right\}$ are the lower and upper bounds on PS_{01}, respectively.

This last estimand is known as the "excess-risk-ratio," and algebra also shows that $\frac{P(Y_1=1)-P(Y_0=1)}{P(Y_1=1)} = 1 - \frac{1}{P(Y_1=1)/P(Y_0=1)}$, where $\frac{P(Y_1=1)}{P(Y_0=1)}$ is the "risk ratio." Thus in this setting, both the "excess-risk-ratio" and the "risk ratio" would be equal across domains. Without monotonicity, the "excess-risk-ratio" is a lower bound on the probability of necessity [137].

Proof. The bounds are obtained by solving a linear optimization problem, as detailed in Appendix 7.4.2. □

Theorem 2 can be better understood as a two-stage process. First, with a little algebra, it is possible to re-express $P^*(Y_1 = 1)$ as a function of PS_{01} alone, resulting in,

$$P^*(Y_1 = 1) = RR \times P^*(Y_0 = 1) + \left(\frac{P(Y_0 = 1) - P^*(Y_0 = 1)}{P(Y_0 = 1)}\right) \text{PS}_{01} \tag{5.2}$$

Where $RR = P(Y_1 = 1)/P(Y_0 = 1)$ denotes the *risk-ratio* obtained in the trial of the source domain Π. The first term of this expression, $RR \times P^*(Y_0 = 1)$, consists of the "naive" prediction for $P^*(Y_1 = 1)$ that one would have obtained by assuming a constant risk ratio across populations. The second term adjusts this naive prediction, by taking into account both the excess risk-ratio of contrasting the baseline mortality between Π and Π^*, as well as the probability of sufficiency shared across environments, PS_{01}.

After this, note that, although the probability of sufficiency PS_{01} in Equation 5.2 cannot be point identified, it can be bounded by (see Appendix 7.4.2 as well as [137])

$$\max\left\{0, \frac{P(Y_1 = 1) - P(Y_0 = 1)}{1 - P(Y_0 = 1)}\right\} \leq \text{PS}_{01} \leq \min\left\{\frac{P(Y_1 = 1)}{1 - P(Y_0 = 1)}, 1\right\} \tag{5.3}$$

Thus, by substituting PS_{01} with its bounds, we obtain the desired bounds for the target quantity $P^*(Y_1 = 1)$.

For instance, in our Russian Roulette example, regardless of whether monotonicity holds, PS_{01} can be bounded by

$$16.7\% \leq \text{PS}_{01} \leq 17.7\%$$

And this assures us that $P^*(Y_1 = 1)$ must lie between,

$$16.8\% \leq P^*(Y_1 = 1) \leq 20.8\%$$

To put it another way, the results of the trial in LA tells us that implementing the policy in

NYC would cause *at least* an increase of $16.8\% - 5\% = 11.8\%$ and *at most* an increase of $20.8\% - 5\% = 15.8\%$ in mortality. Note that, here, substituting the lower bound for PS_{01} (16.7%) actually translates to the *upper bound* for $P^*(Y_1 = 1)$ (20.8%). This happens because the baseline risk in the target population Π^* is *higher* than that of the source population Π, and thus the adjustment due to PS_{01}, in Equation 5.2, is negative.

These considerations naturally lead to the question: in general, how informative are the bounds on $P^*(Y_1 = 1)$? It turns out that the width of the bounds have a simple characterization. Consider the case in which the bounds for PS_{01} are not zero nor one. Now let $P^{*U}(Y_1 = 1)$ and $P^{*L}(Y_1 = 1)$ denote the upper and lower bound on $P^*(Y_1 = 1)$, respectively. After some algebra, it is possible to show that (see Appendix 7.4.2),

$$P^{*U}(Y_1 = 1) - P^{*L}(Y_1 = 1) = \frac{|P(Y_0 = 1) - P^*(Y_0 = 1)|}{1 - P(Y_0 = 1)} \tag{5.4}$$

That is, in this setting, the width of the bounds depends on the baseline risks $P(Y_0 = 1)$ and $P^*(Y_0 = 1)$ alone. Moreover, even if the bounds for PS_{01} happen to be "wide," if the baseline risks are close enough across populations, the bounds for $P^*(Y_1 = 1)$ can still be "narrow." In Section 5.4 we illustrate this fact with a real data example in which the bounds are narrow enough to imply a positive effect of the treatment.

5.3.4 Identification with trials from multiple source domains

In Theorem 2 we learned that the existence of experimental data from *one* source population leads to bounds on the transported causal effect of the target population, although it is not enough for its point identification. Surprisingly, however, if we can obtain experimental data from an additional source population, this suffices to change the picture. With *two* source trials, it is possible to obtain a point estimate for the probabilities of sufficiency, and, consequently, for $P^*(Y_1 = 1)$ without invoking monotonicity, nor any further assumptions beyond $Y_1 \perp\!\!\!\perp S \mid Y_0$. Moreover, multiple source trials entail strong testable implications that can be used to *falsify* this “cross-world” assumption.[11]

To illustrate, consider our Russian Roulette example, and suppose we learn that the city of Chicago has also performed an RCT. In that trial, 25% of those assigned to play the game died, in contrast to 10% of those not assigned to play. If the selection diagram contrasting NYC with Chicago is the same as that of Figure 5.3, we can combine the results from LA and Chicago to estimate the probabilities of sufficiency shared across cities. By the law of total probability, expand the expression for $P(Y_1 = 1)$, both for LA and Chicago, to obtain a system of two equations and two unknowns:

$$\text{(LA Equation):} \qquad 0.175 = (1 - \text{PS}_{10}) \times 0.01 + \text{PS}_{01} \times 0.99 \tag{5.5}$$

$$\text{(Chicago Equation):} \qquad 0.250 = (1 - \text{PS}_{10}) \times 0.10 + \text{PS}_{01} \times 0.90 \tag{5.6}$$

This system can then be solved for PS_{10} and PS_{01}

$$\text{PS}_{10} = 0, \qquad \text{PS}_{01} = 1/6$$

Put differently, the *only* values for PS_{10} and PS_{01} that are compatible with the observed data from *both* trials (LA and Chicago) are that: (i) the “treatment” cannot save anyone from dying; and, that (ii) the treatment kills 1/6 of those who would not have died otherwise. These are the same numeric values as before, but with an important difference—we did not assume monotonicity to obtain point identification; instead, we learned *from the data* that the treatment effect must be monotonic. Once we have these numbers, we can use the same strategy as before to predict the causal effect in NYC, which amounts to, again, 20.8%.

Furthermore, since PS_{10} and PS_{01} must be valid probabilities, not all observed values are compatible with the assumption that $Y_1 \perp\!\!\!\perp S \mid Y_0$. For instance, suppose that instead of 10%, the observed baseline mortality rate in Chicago were 5%. This would imply the impossible value $\text{PS}_{10} = -1.03$, thus *falsifying* the assumption of invariance across domains.

[11] Similar observations regarding testable implications when combining information from multiple studies have also been made in [71], [95] and [44].

It is also easy to see that with three or more source domains we obtain over-identification, since each population pair implies different estimates for PS_{10} and PS_{01}. If those estimates are discordant, this calls into question the assumption of $Y_1 \perp\!\!\!\perp S \mid Y_0$. These results are somewhat reassuring. They tell us that, despite its "cross-world" nature, the assumption of invariance of probabilities of causation across domains may have strong testable implications, and can thus be subjected to empirical scrutiny.

We formalize the previous considerations with the next two theorems.

Theorem 3. *Consider two source domains Π^a and Π^b. Let the probabilities of sufficiency be the same across the two populations, that is, $PS^a_{01} = PS^b_{01} = PS_{01}$ and $PS^a_{10} = PS^b_{10} = PS_{10}$. Then,*

$$
\begin{aligned}
PS_{10} &= 1 - \frac{P^a_{11}P^b_{00} - P^b_{11}P^a_{00}}{P^a_{01}P^b_{00} - P^b_{01}P^a_{00}} \\
PS_{01} &= \frac{P^b_{11}P^a_{01} - P^a_{11}P^b_{01}}{P^a_{01}P^b_{00} - P^b_{01}P^a_{00}}
\end{aligned}
$$

Where $P^a_{ij} := P^a(Y_i = j)$ and $P^b_{ij} := P^b(Y_i = j)$. Moreover, the experimental probabilities of necessity, and probability of necessity and sufficiency [137] of both populations are also identifiable from experimental data of Π^a and Π^b.

Proof. As explained in the text, we can use the law of total probability for each domain to obtain two linear equations with two unknowns, PS_{01} and PS_{10}. We can thus (generically) solve the system of equations for those quantities. Interestingly, in this setting, not only the probabilities of sufficiency, but *all* remaining probabilities of causation (as discussed in [137]), are also identifiable. See details in Appendix 7.4.2. □

Next, the causal effect for a target population Π^* can be transported by appealing again to the law of total probability.

Theorem 4. *Consider two source domains Π^a, Π^b, and a target domain Π^*. Let the probabilities of sufficiency be the same across populations, that is, $PS^a_{01} = PS^b_{01} = PS^*_{01}$ and*

$PS^a_{10} = PS^b_{10} = PS^*_{10}$. *Then, the causal effect* P^*_{11} *in* Π^* *is given by,*

$$P^*_{11} = \frac{P^a_{11}P^b_{00} - P^b_{11}P^a_{00}}{P^a_{01}P^b_{00} - P^b_{01}P^a_{00}} \times P^*_{01} + \frac{P^b_{11}P^a_{01} - P^a_{11}P^b_{01}}{P^a_{01}P^b_{00} - P^b_{01}P^a_{00}} \times P^*_{00}$$

5.4 A Bayesian approach to estimation

The previous results focused on *identification*, that is, they are "asymptotic," and assume that the measured quantities are representative of their corresponding quantities in the population. In practice, however, researchers need to take sampling uncertainty into account. In this section, we describe a Bayesian framework that practitioners can easily put to use for finite sample inference. A Bayesian approach is especially suited for this setting—when the target quantity $P^*(Y_1 = 1)$ is not identifiable from the data alone, preference for any value of the parameter within the identified bounds must rely on prior knowledge.

5.4.1 Model specification

The Bayesian specification of our model can be simplified if we use *counts*. For the source population Π, let n_0 denote the *sum* of individuals with $Y = 1$ in the control group, and let n_1 denote the *sum* of individuals with $Y = 1$ in the treatment group. Likewise, let n^*_0 and n^*_1 denote those quantities for the target population Π^*. Note that n^*_1 is not observed, since the target population is under the "no-treatment" regime.

Now let us use the same notation of Theorem 2 to denote population parameters, that is: $P_{11} := P(Y_1 = 1)$, $P_{01} := P(Y_0 = 1)$, $P^*_{01} := P^*(Y_0 = 1)$, $P^*_{11} := P^*(Y_1 = 1)$. Given that the outcome variable Y is binary, the sum of individuals with $Y = 1$ follows a binomial distribution, and we can write the model for the observed data $\mathcal{D} = \{n_0, n_1, n^*_0\}$ as,

$$n_0 \sim \text{Binomial}(N_0, P_{01}) \tag{5.7}$$

$$n_1 \sim \text{Binomial}(N_1, P_{11}) \tag{5.8}$$

$$n^*_0 \sim \text{Binomial}(N^*_0, P^*_{01}) \tag{5.9}$$

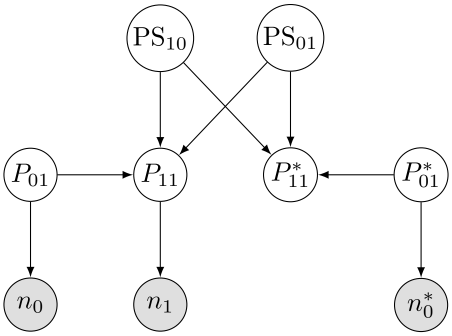


Figure 5.4: Probabilistic graphical model for Bayesian inference when the quantity of interest is P^*_{11}. Gray nodes (n_0, n_1, n^*_0) denote observed variables. White notes denote latent parameters (P_{01}, P_{11}, PS_{10}, PS_{01}, P^*_{11}, P^*_{01}). Note that P_{11} and P^*_{11} share the parameters PS_{10} and PS_{01}, which are invariant across populations.

where N_0 denotes the total number of individuals in the control arm, and N_1 the total number of individuals in the treatment arm of the trial in the source population; N^*_0 denotes the total sample size of the target population (which is under the no-treatment regime). We treat N_0, N_1 and N^*_0 as *known* fixed quantities. Note the observed data depends *only* on the parameters P_{01}, P_{11} and P^*_{01}.

We now need to specify the prior distribution of the parameters and the target quantities of interest. Here we describe two general alternatives, depending on whether the researcher is interested in making inferences directly on P^*_{11} (which in general will not be identified from the data), or on its bounds (which are identified)—we believe these two approaches are complementary, and we encourage investigators to explore both options (see also 116, 70, 128).

Inference on P^*_{11}. As discussed in the previous section, we have that P_{11} is a deterministic function of PS_{10}, PS_{01} and P_{01}, that is, $P_{11} = (1 - \text{PS}_{10})P_{01} + \text{PS}_{01}(1 - P_{01})$. Therefore, we need only to specify priors for the parameters P_{01}, P^*_{01}, PS_{10} and PS_{01}. For example, an "uninformative" (or "flat") prior consists of a uniform distribution over 0 and 1 for all parameters. Another option is to choose a prior that incorporates the assumption of monotonicity, by setting a point mass on $\text{PS}_{10} = 0$. Users have the flexibility of picking anything in between, such as setting a prior that puts most, but not all, of the mass on $\text{PS}_{10} = 0$, for instance. The target of inference is the *posterior distribution* of P^*_{11}, which is, again, a transformation of the parameters P^*_{01}, PS_{10} and PS_{01},

$$P^*_{11} = (1 - \mathrm{PS}_{10})P^*_{01} + \mathrm{PS}_{01}(1 - P^*_{01})$$

As we shall see, with a "flat" prior, as the sample size increases the posterior distribution remains spread on the identified bounds; whereas with a prior that assumes monotonicity the posterior converges to the identified point estimate. Other quantities of interest may be the posterior distribution of certain *effect measures*, such as the risk difference $RD^* = P^*_{11} - P^*_{01}$ or the risk ratio $RR^* = P^*_{11}/P^*_{01}$. Figure 5.4 shows the probabilistic graphical model of this setup, with observed variables in gray, and latent parameters in white. The known fixed parameters N_0, N_1 and N^*_0 are omitted for clarity.

Inference on bounds. When making inferences on P^*_{11} (which is not identified), the shape of its posterior will be dependent on (but not completely determined by) the shape of the prior of the unidentified quantities PS_{01} and PS_{10}, regardless of sample size. For this reason, users may also find useful to perform inference directly on the bounds P^{*L}_{11} and P^{*U}_{11} (which are identified). While the previous framework can still be used for such inferences, we note that, if interest lies on the bounds alone, there is a simpler alternative—as the bounds are functionals of the observed data, inference about P^{*L}_{11} and P^{*U}_{11} only requires priors on the identified parameters P_{01}, P_{11} and P^*_{01} [116, 128].

Sampling. Given the observed data $\mathcal{D}$ and a prior distribution on the parameters, one can obtain the posterior distribution of the target quantities using Gibbs sampling. Here we use the Gibbs sampler JAGS [115]. Extending the model to two (or more) source populations follows the same logic, thus we defer its discussion to Appendix 7.4.4. Next, we demonstrate the method using: (i) simulated data from the Russian Roulette example; and, (ii) real data from trials that investigate the effects of vitamin A supplementation on childhood mortality. Code for replicating all results is also provided in Appendix 7.4.4.

5.4.2 Simulated data example

To illustrate the method, we start by applying our tools to simulated data drawn from a process with the same proportions as the Russian Roulette example, with various sample sizes. We show the posterior distribution of $P^*(Y_1 = 1)$ using both a "flat" prior for all parameters, and a prior assuming monotonicity. The results are shown in Figures 5.5 and 5.6.

Let us start by examining Figure 5.5. Here we set "flat" priors for *all* parameters. Note that, as per Theorem 2, the posterior distribution remains spread in the asymptotic bounds of 16.8% and 20.8% regardless of sample size. Moving to Figure 5.6, we now set a point mass prior on $\text{PS}_{10} = 0$, representing the assumption of monotonicity. The remaining parameters continue to have a "flat" prior. As expected, the posterior distribution now concentrates around 20.8% as the number of cases increases.

5.4.3 Real data example

We now illustrate our method with a real data example. We investigate three experiments designed to determine the effects of vitamin A supplementation on childhood mortality. The first trial was carried out in the Aceh province at the northern tip of Sumatra, Indonesia [132]; the second trial was conducted in the West Java province, in Java, also in Indonesia [102]. Finally, the third trial took place in the district of Sarlahi, Nepal [144]. The results from the studies are shown in Table 5.1. Our exercise in this section consists of using the results of earlier trials, along with the baseline risk of the target population, to predict mortality under treatment in the target population.

Study	Treatment		Control	
	Survived	Total	Survived	Total
Aceh [132]	12,890	12,991	12,079	12,209
West Java [102]	5,589	5,775	5,195	5,445
Sarlahi [144]	14,335	14,487	13,933	14,143

Table 5.1: Observed data for the vitamin A studies.

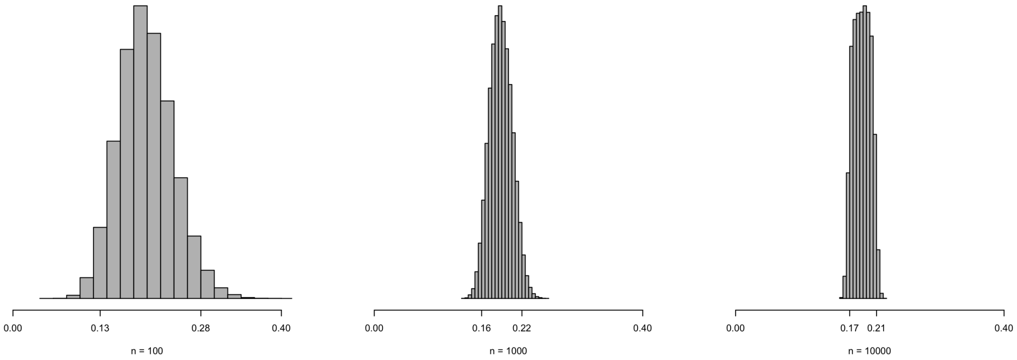


Figure 5.5: Histograms of the posterior samples of $P^*(Y_1 = 1)$ for a simulation of the Russian Roulette data, considering different sample sizes 100, $1,000$ and $10,000$. Here all parameters have a "flat" prior. Note that, as the sample size increases, the posterior distribution does not concentrate on a point; rather, the posterior remains spread on the identified bound of 16.8% to 20.8%, as per Theorem 2.

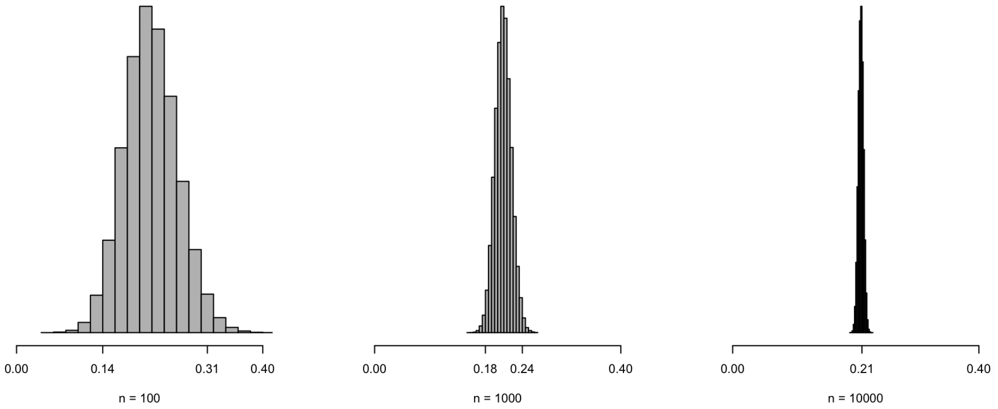


Figure 5.6: Histograms of the posterior samples of $P^*(Y_1 = 1)$ for a simulation of the Russian Roulette data, considering different sample sizes 100, $1,000$ and $10,000$. Here we put a point mass prior on PS_{10}, corresponding to the assumption of monotonicity. The remaining parameters have a "flat" prior. Note that, as the sample size increases, the posterior distribution concentrates on 20.8%, since the parameter is identifiable in this setting.

It is suspected that vitamin A reduces childhood mortality by reducing the incidence, severity or duration of life-threatening diseases such as measles and diarrhoea [144]. As a *first approximation* to this process, we can borrow the same disjunctive model of the previous section. The variables now mean: (i) $Y = 1$ *survival*, and $Y = 0$ death during the trial; (ii) $H = 1$ *absence*, and $H = 0$ presence of severe measles; (iii) $X = 1$ participation in the treatment group (vitamin A supplementation), and $X = 0$ participation in the control group; finally, (iv) B summarizes biological factors that determine the response to treatment ($B = 1$ successful response, $B = 0$ otherwise). Here the monotonicity assumption states that vitamin A supplementation *does not* cause deaths. After presenting the results of our method, we discuss cases under which these assumptions may be violated, thus preventing one from inferring $Y_1 \perp\!\!\!\perp S \mid Y_0$.

Our first task is to use the results of the Aceh trial (Π^A) to predict the effects of the West Java trial (Π^{WJ}). The estimates of the Aceh trial are $\widehat{P}^A(Y_1 = 1) = 0.992$ and $\widehat{P}^A(Y_0 = 1) = 0.989$; whereas the baseline risk in the Java trial is $\widehat{P}^{WJ}(Y_0 = 1) = 0.954$. As expected, note the large discrepancy of baseline risk in both trials, indicating the existence of structural differences in how mortality is determined, and thus forbidding a direct transport of $P^{WJ}(Y_1 = 1)$. Figure 5.7 shows the posterior distribution of $P^{WJ}(Y_1 = 1)$ using both a “flat” prior for all parameters (left), and a prior assuming monotonicity for the effect of vitamin A supplementation (right). In the first case, we obtain a 95% *credible interval* of 0.962 to 0.992 for $P^{WJ}(Y_1 = 1)$, in agreement with the asymptotic bounds of Theorem 2—this shows that, even without assuming monotonicity, the bounds are narrow enough to be consistent with a positive effect of vitamin A supplementation in West Java.[12] When assuming a monotonic effect of vitamin A, we obtain the posterior mean of 0.967 (95% CI 0.956–0.975). In both plots, a red dashed line indicates the actual value observed in the West Java trial, $\widehat{P}^{WJ}(Y_1 = 1) = 0.968$, which is consistent with the predictions of our method.

[12] The 95% credible intervals for the risk difference and risk ratio are 0.008–0.04 and 1.009–1.042, respectively. Alternatively, if one prefers inferences on the bounds, we have 95% credible intervals of: 0.955–0.975 for the lower bound, 0.991–0.994 for the upper bound, and 0.002–0.020 for the lower bound of the risk difference (i.e, $P^{*L}_{11} - P^{*}_{01}$).

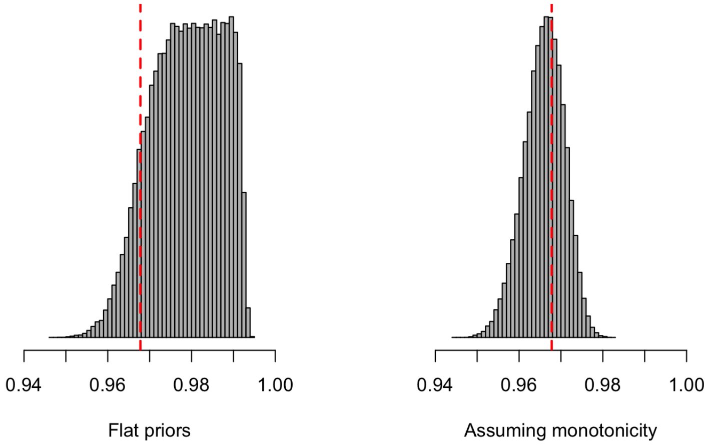


Figure 5.7: Posterior of $P^{WJ}(Y_1 = 1)$ for the West Java trial, using data from the Aceh trial. Left: posterior of $P^{WJ}(Y_1 = 1)$ using "flat" priors. Right: posterior of $P^{WJ}(Y_1 = 1)$ assuming monotonicity. Red dashed lines show the observed value in the West Java trial, $\widehat{P}^{WJ}(Y_1 = 1)$.

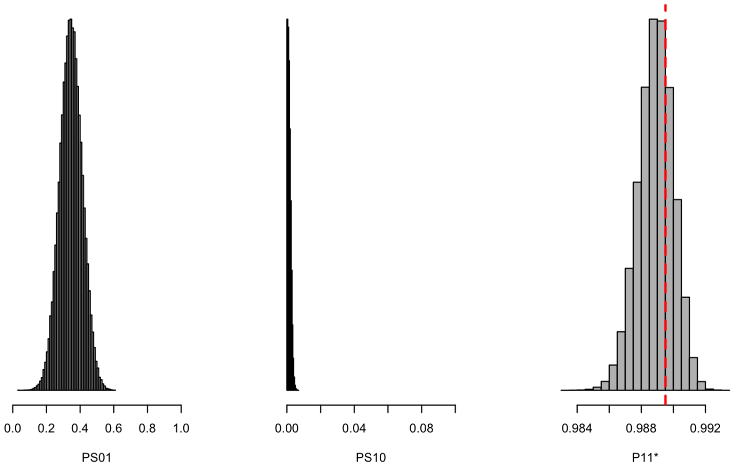


Figure 5.8: From left to right, posterior of PS_{01}, PS_{10} and $P^S(Y_1 = 1)$ using data from *both* the Aceh and West Java trials [132, 102], and using "flat" priors for all parameters. Dashed red line indicates the observed value in the Sarlahi trial, $\widehat{P}^S(Y_1 = 1)$.

Our second task is to use the results of *both* the Aceh (Π^A) and West Java (Π^{WJ}) trials to predict the effects of the Sarlahi trial (Π^S). As per Theorems 3 and 4, in this setting we can identify the probabilities of sufficiency shared across regions, PS_{10} and PS_{01}, as well as the effect in Sarlahi, $P^S(Y_1 = 1)$, *without* assuming monotonicity. The posterior distributions of these three quantities are displayed in Figure 5.8. The posterior mean for PS_{01} is 0.346 (95% CI 0.214–0.478), while the posterior mean for PS_{10} is 0.001 (95% CI 0.000–0.004). This suggests that, in the context of these trials, vitamin A supplementation is sufficient to prevent 21% to 48% of the deaths that would have otherwise occurred without supplementation, while it has no or little side-effects that are sufficient to cause the death of otherwise healthy subjects. Finally, we obtain the posterior mean of 0.989 (95% CI 0.987–0.991) for $P^S(Y_1 = 1)$, consistent with the actual value observed in the Sarlahi trial, $\widehat{P}^S(Y_1 = 1) = 0.989$.

Before moving to the conclusions, let us use this example to make some brief remarks about causal modeling in practice. Note that the working model in this section assumes the only factor causing deaths during the period of the trial can be summarized by H, consisting of diseases which, at least in principle, can be affected by the treatment (e.g, severe measles or diarrhoea). What happens, however, if we augment the model to allow for other causes of deaths unaffected by vitamin A supplementation? It can be shown that this new variable is a common cause of both potential responses, thus creating a colliding path and forbidding the conclusion that $Y_1 \perp\!\!\!\perp S \mid Y_0$.[13] This suggests caution when transporting these results to populations where mortality due to diarrhoea or measles is not predominant.

More generally, while one may summarize the main "identification assumption" for the results in this chapter in terms of the counterfactual independence $Y_1 \perp\!\!\!\perp S \mid Y_0$, note we did not commence the analysis by imposing this or any "identification assumption." Instead, we made an effort to explicate our understanding of the problem directly in a structural model, and the necessary counterfactual independence emerged naturally as a *logical consequence of*

[13] Call these new causes C. The new structural equation for Y now reads $Y = (H \vee (X \wedge B)) \wedge \neg C$. This leads to $Y_0 = H \wedge \neg C$ and $Y_1 = Y_0 \vee (B \wedge \neg C)$. Note this creates the colliding path $S \to H \to Y_0 \leftarrow C \to Y_1$, thus forbidding the conclusion that $Y_1 \perp\!\!\!\perp S \mid Y_0$, even when there is no selection node pointing directly to C. For another illustration of when collider bias may arise, see Appendix 7.4.3.

the structure. This is an important part of the process. If some of those modeling assumptions happen to be challenged, as they often are in practical settings (e.g, unobserved confounding between H and B), we should refrain from positing that $Y_1 \perp\!\!\!\perp S \mid Y_0$ and the model both warns us of possible threats, as well as helps us in finding alternative solutions.[14]

5.5 Conclusions

This chapter showed how two apparently separate areas of causal inference research—the generalization of causal effects across populations [113, 16, 78] and the identification of "causes of effects" [108, 137, 111, 112]—can be merged for mutual benefit, unveiling important results in both areas.

The first lesson that emerges from this combined analysis is that certain functional constraints may entail the invariance of probabilities of causation across domains, which can then be used as instruments to license generalization. This may occur when the outcome is a product of several independent processes, only some of which are carriers of disparities, and when the outcome produced under the "no-treatment" condition is sufficient to block these sources of disparity. These functional constraints may enable the identification, or at least the bounding of the target effect in settings where non-parametric generalization is otherwise impossible.

A second lesson that surfaces from our investigation is that, whenever experimental data from multiple sites are available, these may lead to the point identification of probabilities of causation. These counterfactual probabilities can be the targets of investigations in public health, legal settings, and the production of explanations [101, 111, 112]. For example, drugs with a positive average treatment effect may still kill individuals who would have otherwise survived—being able to quantify the percentage of individuals that are saved or harmed by the treatment has important implications in many public health applications.

[14] For example, a sensitivity analysis might still be possible, and one could investigate how big a departure from the original model assumptions would be necessary to invalidate the main conclusions.

The development of tools for automating the types of analyses presented here, paralleling those available for non-parametric models, is a challenging topic for future work. As we have seen, determining the invariance of probabilities of causation requires additional constraints beyond the standard non-parametric model; some recent developments, such as algorithms for handling context-specific independencies for causal identification [139], may provide the initial steps towards this undertaking.

CHAPTER 6

Conclusion

In this dissertation we developed new theory, methods, and software for drawing causal inferences under more flexible and realistic settings. More specifically, Chapters 2 and 3 developed a novel powerful, yet simple, suite of sensitivity analysis tools for popular methods, such as confounding adjustment and instrumental variables; Chapter 4 devised the first systematic, algorithmic approach to sensitivity analysis for *arbitrary* linear structural causal models, subsuming many previous canonical results of prior literature; and Chapter 5 derived novel (partial) identification results both for the generalization of causal effects across populations as well as for the identification of "causes of effects." Each of these projects represent an ongoing research agenda, with promising directions for future work.

In particular, I believe the methods developed in Chapters 2 and 3 have the potential to quickly become the de-facto standard for sensitivity analysis, and soon be ubiquitous in applied papers. An extension of the methods presented in Chapter 3 to the area of Mendelian Randomization (using genetic variants as instrumental variables) is already under work [38]. Short to medium term goals in this area should include the development of a suite of sensitivity analysis tools for panel data methods and regression discontinuity designs, as well as extending these results to more flexible semiparametric models. This would cover the bulk of the main methods that are currently widely used in applied work, making sensitivity analysis easy to peform, routine and standard practice across the applied sciences.

The overarching theme around the results of Chapter 4 is to devise new ways to represent soft constraints on the data generating process (or alternative constraints currently neglected by traditional theory), along with tools to *systematically* derive (partial) identification results leveraging such constraints. I believe this will be an indispensable part for a modern, flexible and trustworthy approach to causal inference, that allows for modeling assumptions that better match the researcher's domain knowledge. Recent explorations in this direction include incorporating arbitrary *relative constraints* among causal effects in linear structural models [147], leading to interesting new identification results (such as generalizing the well known differences-in-differences method), or to the ability to systematically leverage knowledge of "variable importance" for benchmarking in sensitivity analysis (as in Section 2.4.4, for instance). Exploiting non-zero constraints on path-specific effects has also led to a new state-of-the-art algorithm for traditional linear identification itself [89]. All these results, however, still rely on the strong assumption of linearity for all variables of the system, and an important direction for future endeavours is to obtain similar results for non-parametric models.

Finally, extensions on the work of generalizability and probabilities of causation of Chapter 5 include a refined taxonomy of causal estimands for these counterfactual quantities, improved bounds from multiple domains, without requiring strict equality of probabilities of causation, as well as a more robust Bayesian workflow for inference with finite samples under partial identification.

CHAPTER 7

Appendices

7.1 Appendix for Chapter 2

7.1.1 Simple measures for routine reporting

7.1.1.1 Preliminaries

For any univariate regression, recall $R^2 = t^2/(t^2 + \text{df})$, $t^2 = \left(\frac{R^2}{1-R^2}\right)\text{df}$, and $f^2 = \frac{R^2}{1-R^2} = \frac{t^2}{\text{df}}$, where df is the regression's degrees of freedom. Repeating the partialling out procedure to allow for covariates, the partial R^2 of any covariate can be written in terms of its coefficient's t statistic and *vice-versa*. For instance, the partial R^2 of the confounder with the treatment, conditional on $\boldsymbol{X}$, can be written as

$$R^2_{D\sim Z|\boldsymbol{X}} = \frac{t^2_{\hat{\delta}}}{t^2_{\hat{\delta}} + \text{df}}. \tag{7.1}$$

Analogously,

$$f^2_{D\sim Z|\boldsymbol{X}} = \frac{R^2_{D\sim Z|\boldsymbol{X}}}{1 - R^2_{D\sim Z|\boldsymbol{X}}} = \frac{t^2_{\hat{\delta}}}{\text{df}}. \tag{7.2}$$

Where $\hat{\delta}$ is the coefficient of the regression equation $Z = \hat{\delta}D + \boldsymbol{X}\hat{\psi} + \hat{\varepsilon}_Z$, and $t_{\hat{\delta}}$ is the t-value corresponding to $\hat{\delta}$.

7.1.1.2 General strength of a confounder

Consider a confounder strong enough to change the estimated treatment effect by $(100 \times q^*)\%$. This means that $|\widehat{\text{bias}}| = q^*|\hat{\tau}_{\text{res}}|$. Hence, by equation 2.13 we have that

$$q^*|\hat{\tau}_{\text{res}}| = \sqrt{\frac{R^2_{Y\sim Z|D,\boldsymbol{X}}\ R^2_{D\sim Z|\boldsymbol{X}}}{1 - R^2_{D\sim Z|\boldsymbol{X}}}\widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\text{df}}}. \tag{7.3}$$

Dividing both sides by $\widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\text{df}}$ and noting $\frac{|\hat{\tau}_{\text{res}}|}{\widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\text{df}}} = \frac{|t_{\hat{\tau}_{\text{res}}}|}{\sqrt{\text{df}}} = f_{Y\sim D|\boldsymbol{X}}$, we obtain

$$\begin{aligned} q^*|f_{Y\sim D|\boldsymbol{X}}| &= \sqrt{\frac{R^2_{Y\sim Z|D,\boldsymbol{X}}\ R^2_{D\sim Z|\boldsymbol{X}}}{1 - R^2_{D\sim Z|\boldsymbol{X}}}} && (7.4)\\ &= |R_{Y\sim Z|D,\boldsymbol{X}} \times f_{D\sim Z|\boldsymbol{X}}| && (7.5)\\ &= \text{BF}. && (7.6) \end{aligned}$$

That is, to bring the estimated effect down by $(100 \times q^*)\%$, the bias factor (BF) of the confounder $\left(R_{Y\sim Z|D,\boldsymbol{X}} f_{D\sim Z|\boldsymbol{X}}\right)$ has to equal q^* times the partial f of the treatment with the outcome.

7.1.1.3 Extreme sensitivity scenarios

Considering the extreme case scenario where the confounders explain all the residual variance of the outcome, that is, $R^2_{Y\sim Z|D,\boldsymbol{X}} = 1$, a confounder strong enough to bring down the estimated effect to zero (that is $q^* = 1$) would need to satisfy $f^2_{Y\sim D|\boldsymbol{X}} = f^2_{D\sim Z|\boldsymbol{X}}$ which implies $R^2_{Y\sim D|\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}}$. This shows the partial R^2 of the treatment with the outcome is itself a measure of an extreme-scenario sensitivity analysis.

7.1.2 The Robustness Value (RV)

Now consider a confounder with $R^2_{Y\sim Z|D,\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}} = \mathrm{RV}_{q^*}$. Rearranging terms and squaring Equation 7.4, one obtains

$$\mathrm{RV}^2_{q^*} + f^2_{q^*}\mathrm{RV}_{q^*} - f^2_{q^*} = 0, \tag{7.7}$$

where $f_{q^*} := q^*|f_{Y\sim D|\boldsymbol{X}}|$. Solving the quadratic equation for RV_{q^*},

$$\mathrm{RV}_{q^*} = \frac{1}{2}\left(\sqrt{f^4_{q^*} + 4f^2_{q^*}} - f^2_{q^*}\right) \tag{7.8}$$

gives us the equation for the robustness value for the point estimate. Note that, since the derivative of the bias with respect to both sensitivity parameters is positive, any confounder with both associations below RV_{q^*} is not strong enough to bring about a relative bias of q^*.

RV for t-values, or lower and upper bounds of confidence intervals

Imagine the researcher wants to know how strong a confounder would need to be for a $100(1-\alpha)\%$ confidence interval to include a change of $(100 \times q^*)\%$ of the treatment estimate. Consider again a confounder with equal association with the treatment and the outcome, $R^2_{Y\sim Z|D,\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}} = \mathrm{RV}_{q^*,\alpha}$. By Equation 2.13,

$$|\hat{\tau}| = |\hat{\tau}_{\text{res}}| - \widehat{\text{se}}(\hat{\tau}_{\text{res}})\frac{\mathrm{RV}_{q^*,\alpha}}{\sqrt{1-\mathrm{RV}_{q^*,\alpha}}}\sqrt{\text{df}}, \tag{7.9}$$

where we are assuming the bias reduces the absolute value of the estimated effect. For the opposite direction the subtraction would be changed to addition. Further, for any confounder with equal association with the treatment and the outcome, Equation 2.12 for the adjusted standard error simplifies to

$$\widehat{\text{se}}(\hat{\tau}) = \widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\frac{\text{df}}{\text{df}-1}}. \tag{7.10}$$

Let $|t^*_{\alpha,\text{df}-1}|$ denote the t-value threshold for a t-test with significance level of α and $\text{df}-1$ degrees of freedom, and define $f^*_{\alpha,\text{df}-1} := |t^*_{\alpha,\text{df}-1}|/\sqrt{\text{df}-1}$. Now note that, for the adjusted t-test to not reject the hypothesis $H_0 : \tau = (1-q)\hat{\tau}_{\text{res}}$, we must have

$$\begin{aligned} |t^*_{\alpha,\text{df}-1}| &\geq \frac{|\hat{\tau}| - (1-q^*)|\hat{\tau}_{\text{res}}|}{\widehat{\text{se}}(\hat{\tau})} && (7.11) \\ &\geq \frac{q^*|\hat{\tau}_{\text{res}}| - \widehat{\text{se}}(\hat{\tau}_{\text{res}})\frac{\text{RV}_{q^*,\alpha}}{\sqrt{1-\text{RV}_{q^*,\alpha}}}\sqrt{\text{df}}}{\widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\frac{\text{df}}{\text{df}-1}}} && (7.12) \\ &\geq \left(q^*|f_{Y\sim D|\boldsymbol{X}}| - \frac{\text{RV}_{q^*,\alpha}}{\sqrt{1-\text{RV}_{q^*,\alpha}}}\right)\sqrt{\text{df}-1}. && (7.13) \end{aligned}$$

Divide by $\sqrt{\text{df}-1}$ and rearrange terms to obtain,

$$\frac{\text{RV}_{q^*,\alpha}}{\sqrt{1-\text{RV}_{q^*,\alpha}}} \geq f_{q^*} - f^*_{\alpha,\text{df}-1} = f_{q^*,\alpha}, \tag{7.14}$$

where we define $f_{q^*,\alpha} := f_{q^*} - f^*_{\alpha,\text{df}-1}$. Our goal is to find the minimal strength of the confounder $\text{RV}_{q^*,\alpha}$ (which must be positive) such that this inequality holds. Thus, we have two cases. If $f_{q^*,\alpha} < 0$, then trivially $\text{RV}_{q^*,\alpha} = 0$. This happens when an inclusion of a covariate with zero predictive power would be enough not to reject the null hypothesis, either because the t-value is already low enough, or because it becomes low enough after adjusting for the loss in degrees of freedom.

Now consider the case where $f_{q^*,\alpha} > 0$, which means the minimum will happen in the equality. Rearrange terms and square to obtain,

$$\text{RV}^2_{q^*,\alpha} + f^2_{q^*,\alpha}\text{RV}_{q^*,\alpha} - f^2_{q^*,\alpha} = 0. \tag{7.15}$$

Solving the quadratic equation for $\text{RV}_{q^*,\alpha}$ gives us the robustness value for a reduction of $(100 \times q^*)\%$ to not be rejected at the significance level α,

$$\text{RV}_{q^*,\alpha} = \frac{1}{2}\left(\sqrt{f^4_{q^*,\alpha} + 4f^2_{q^*,\alpha}} - f^2_{q^*,\alpha}\right). \tag{7.16}$$

Note that, due to the variance reduction factor of Equation 2.15, it could be the case that increasing the sensitivity parameter $R^2_{Y\sim Z|D,\boldsymbol{X}}$ helps with statistical significance. When this happens, there can exist a set of confounders with lower $R^2_{Y\sim Z|D,\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$ than $\text{RV}_{q^*,\alpha}$ able to drive the t-statistic below significance. To check for such cases, we need to verify whether the derivative of the adjusted t-value with respect to $R^2_{Y\sim Z|D,\boldsymbol{X}}$ is negative (the derivative with respect to $R^2_{D\sim Z|\boldsymbol{X}}$ is always negative). The t-value for $\hat{\tau}$ for testing the null hypothesis $H_0 : \tau = (1-q)\hat{\tau}_{\text{res}}$ can be written as,

$$t_{\hat{\tau},q^*} = \frac{\hat{\tau} - (1-q^*)\hat{\tau}_{\text{res}}}{\widehat{\text{se}}(\hat{\tau})} = \frac{f_{q^*}\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}}} - \sqrt{R^2_{Y\sim Z|D,\boldsymbol{X}}}\sqrt{R^2_{D\sim Z|\boldsymbol{X}}}}{1-\sqrt{R^2_{Y\sim Z|D,\boldsymbol{X}}}} \times \sqrt{\text{df}-1} \tag{7.17}$$

Dividing by $\sqrt{\text{df}-1}$ and taking the derivative with respect to $R^2_{Y\sim Z|D,\boldsymbol{X}}$ gives us,

$$\frac{\partial t_{\hat{\tau},q^*}}{\partial R^2_{Y\sim Z|D,\boldsymbol{X}}} = \frac{f_{q^*}\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}}}\sqrt{R^2_{Y\sim Z|D,\boldsymbol{X}}} - \sqrt{R^2_{D\sim Z|\boldsymbol{X}}}}{2\sqrt{R^2_{Y\sim Z|D,\boldsymbol{X}}}(1-R^2_{Y\sim Z|D,\boldsymbol{X}})^{3/2}} \tag{7.18}$$

Equation 7.18 is negative when the numerator is less than zero, that is, when

$$\frac{R^2_{D\sim Z|\boldsymbol{X}}}{(1-R^2_{D\sim Z|\boldsymbol{X}})R^2_{Y\sim Z|D,\boldsymbol{X}}} > f^2_{q^*} \tag{7.19}$$

For the point of equal association, $R^2_{Y\sim Z|D,\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}} = \text{RV}_{q^*,\alpha}$, the condition in Equation 7.19 simplifies to $\text{RV}_{q^*,\alpha} > 1 - 1/f^2_{q^*}$. Note that, since $RV \geq 0$ this condition will often hold—for instance, for $q^* = 1$, whenever the partial R^2 of the treatment with the outcome is less or equal to 50%, the first order condition is guaranteed to hold.

When condition 7.19 does not hold, Equation 7.16 is still a useful and meaningful reference point of a specific contour line. However, one may want to alternatively define the $\text{RV}_{q^*,\alpha}$ as the maximum bound on both coordinates such that any confounder with (both) associations below that bound cannot bring the t-value below the chosen critical level. In that case, given a bound of $\text{RV}_{q^*,\alpha}$ on both coordinates, we can solve the following constrained minimization problem,

$$\min_{R^2_{Y\sim Z|D,\boldsymbol{X}},R^2_{D\sim Z|\boldsymbol{X}}} t_{\hat{\tau},q} \;\; \text{s.t.} \;\; R^2_{Y\sim Z|D,\boldsymbol{X}} \leq \text{RV}_{q^*,\alpha} \text{ and } R^2_{D\sim Z|\boldsymbol{X}} \leq \text{RV}_{q^*,\alpha} \tag{7.20}$$

Since the derivative of the adjusted t-value with respect to $R^2_{D\sim Z|\boldsymbol{X}}$ is always negative, the optimum $R^2_{D\sim Z|\boldsymbol{X}}$ always reaches the bound. Next we have two cases: when (i) the derivative of the solution with respect to $R^2_{Y\sim Z|D,\boldsymbol{X}}$ is negative, this means the optimum for both arguments reach the bound, and solving for a specific t-value threshold gives $\text{RV}_{q^*,\alpha}$ as before (Equation 7.16); when (ii) the derivative of the solution with respect to $R^2_{Y\sim Z|D,\boldsymbol{X}}$ is zero, then the optimal $R^2_{Y\sim Z|D,\boldsymbol{X}}$ is an interior point, which by Equation 7.19 equals $R^2_{Y\sim Z|D,\boldsymbol{X}} = \text{RV}_{q^*,\alpha}/((1-\text{RV}_{q^*,\alpha})f^2_{q^*})$. Solving for a specific t-value threshold gives us the bound,

$$\text{RV}_{q^*,\alpha} = \frac{f^2_{q^*} - f^{*2}_{\alpha,\text{df}-1}}{1+f^2_{q^*}} \tag{7.21}$$

Finally, note that if one picks the threshold $|t^*_{\alpha,\text{df}-1}| = 0$ then $\text{RV}_{q^*,\alpha}$ trivially reduces to RV^*_q. Also note that, for fixed $|t^*_{\alpha,\text{df}-1}|$, when $df \rightarrow \infty$ we have that $\text{RV}_{q^*,\alpha} \rightarrow \text{RV}_{q^*}$, since standard errors become irrelevant when compared to the bias of the point estimate.

7.1.2.1 *Impact thresholds* [58] for non-zero null hypothesis

In Section 2.6.1 we showed that a confounder's *impact*, as defined in [58], does not fully characterize the minimal strength of confounding necessary to bring about a certain amount of bias in the regression coefficient, except when the relative bias is unity (that is, when the null hypothesis of interest is *zero*). Thus, the *impact thresholds* obtained in [58] under the null of zero (in which case $R_{Y\sim Z|\boldsymbol{X}} = R_{D\sim Z|\boldsymbol{X}}$) cannot be immediately generalized to non-zero null hypotheses. Here we provide a simple illustrative numerical example. Consider the case with no observed covariates $\boldsymbol{X}$, a single unobserved confounder Z, all variables standardized to mean zero and unit variance and a sample of size $1,000$. Suppose $\hat{\tau}_{\text{res}} = 0.5$,

$\widehat{\text{se}}(\hat{\tau}_{\text{res}}) = 0.0274$ and that we want to learn the minimal strength of Z necessary to bring this estimate to $\hat{\tau} = -0.5$ (a relative bias of 2). Solving the bias equation for the case where $R_{Y\sim Z|\boldsymbol{X}} = R_{D\sim Z|\boldsymbol{X}}$ one would obtain an impact threshold of 2/3. However, this is not the minimal *impact* that would make $\hat{\tau} = -0.5$. As a counterexample, a confounder with an *impact* as low as 0.51 is sufficient to bring about a change of this magnitude, with $R_{Y\sim Z|\boldsymbol{X}} = 0.515$ and $R_{D\sim Z|\boldsymbol{X}} = 0.99$.

7.1.3 Formal benchmark bounds

Suppose the researcher has substantive knowledge that certain covariates are "the most important predictors of the outcome" and other covariates "the most important predictors of the treatment assignment." Imagine, also, that the researcher is willing to defend the claim that the unobserved confounder Z is not "as strong" as those covariates.

In order to use this information for bounding the strength of the confounder Z, we need to give it an operational meaning. We operationalize these types of claim as comparisons of the explanatory power of the confounder vis-a-vis the explanatory power of the observed covariates. Mathematically, we can quantify these comparisons using total or partial R^2 measures. Here we will assume that $Z \perp \boldsymbol{X}$ or, equivalently, that the following analysis applies to the part of Z not linearly explained by covariates $\boldsymbol{X}$.

7.1.3.1 Comparing the total R^2 of covariates with the total R^2 of the confounder

Although in the text we use the bounds by comparing partial R^2 measures, perhaps the simplest derivation is the comparison of the total R^2 of observed covariates with the total R^2 of the unobserved confounder Z. Consider an example in which the observed covariate X_j is assumed to be an important predictor of the treatment assignment D. If the researcher believes the correlation of X_j with D to be stronger than the correlation of Z with D, this implies,

$$R^2_{D\sim Z} < R^2_{D\sim X_j}. \tag{7.22}$$

We could use the same argument for comparing $R^2_{Y\sim Z}$ with $R^2_{Y\sim X_j}$. As it happens, such claims are sufficient to bound the sensitivity parameters. Let us generalize this notion by defining,

$$k_D := \frac{R^2_{D\sim Z}}{R^2_{D\sim X_j}}, \qquad k_Y := \frac{R^2_{Y\sim Z}}{R^2_{Y\sim X_j}}. \tag{7.23}$$

That is, k_D and k_Y measure how the correlation of Z, with D and Y, compares to the correlation of X_j with those same variables. Our goal here is to re-express both sensitivity parameters as a function of k_D and k_Y. Since $Z \perp \boldsymbol{X}$, we have that

$$R^2_{D\sim Z+\boldsymbol{X}} = R^2_{D\sim Z} + R^2_{D\sim \boldsymbol{X}} = k_D R^2_{D\sim X_j} + R^2_{D\sim \boldsymbol{X}} \tag{7.24}$$

$$R^2_{Y\sim Z+\boldsymbol{X}} = R^2_{Y\sim Z} + R^2_{Y\sim \boldsymbol{X}} = k_Y R^2_{Y\sim X_j} + R^2_{Y\sim \boldsymbol{X}}. \tag{7.25}$$

Now we can trivially re-express $R^2_{D\sim Z|\boldsymbol{X}}$ as function of k_D,

$$R^2_{D\sim Z|\boldsymbol{X}} = \frac{R^2_{D\sim Z+\boldsymbol{X}} - R^2_{D\sim \boldsymbol{X}}}{1 - R^2_{D\sim \boldsymbol{X}}} \tag{7.26}$$

$$= k_D \left(\frac{R^2_{D\sim X_j}}{1 - R^2_{D\sim \boldsymbol{X}}} \right). \tag{7.27}$$

Analogous result holds for $R^2_{Y\sim Z|\boldsymbol{X}}$. What remains is to re-express $R^2_{Y\sim Z|D,\boldsymbol{X}}$. Using the standard recursive definition of partial correlations, we know that

$$\left| R_{Y\sim Z|\boldsymbol{X},D} \right| = \frac{\left| R_{Y\sim Z|\boldsymbol{X}} - R_{Y\sim D|\boldsymbol{X}} R_{D\sim Z|\boldsymbol{X}} \right|}{\sqrt{1 - R^2_{Y\sim D|\boldsymbol{X}}}\sqrt{1 - R^2_{D\sim Z|\boldsymbol{X}}}}. \tag{7.28}$$

The only two terms of the RHS including the confounder, $R_{Y\sim Z|\boldsymbol{X}}$ and $R_{D\sim Z|\boldsymbol{X}}$, have been re-expressed as a function of k_D and k_Y above. We now show how to determine the sign of the correlations, by considering the direction of the strengths of the confounder act towards hurting our preferred hypothesis.

Let us assume the confounder acts towards reducing the absolute value of the effect size. If the effect size is positive ($R_{Y\sim D|\boldsymbol{X}} > 0$), this means $R_{Y\sim Z|D,\boldsymbol{X}}$ and $R_{D\sim Z|\boldsymbol{X}}$ must have the same signs. Consider, first, $R_{Y\sim Z|\boldsymbol{X},D} < 0$ and $R_{D\sim Z|\boldsymbol{X}} < 0$. This implies $R_{Y\sim Z|\boldsymbol{X}} < 0$, which means we are reducing the absolute value of $R_{Y\sim Z|\boldsymbol{X}}$. Now consider $R_{Y\sim Z|D,\boldsymbol{X}} > 0$ and $R_{D\sim Z|\boldsymbol{X}} > 0$. This implies $R_{Y\sim Z|\boldsymbol{X}} > 0$, which, again, means we are reducing the absolute value of $R_{Y\sim Z|\boldsymbol{X}}$. If the effect size is negative ($R_{Y\sim D|\boldsymbol{X}} < 0$), this now would mean that $R_{Y\sim Z|D,\boldsymbol{X}}$ and $R_{D\sim Z|\boldsymbol{X}}$ must have the opposite signs, and applying the previous arguments, we reach the same conclusion that we will be reducing the absolute value of $R_{Y\sim Z|\boldsymbol{X}}$.

Therefore, considering that the confounder acts towards *reducing* the magnitude of the estimate towards zero, we have that,

$$\left|R_{Y\sim Z|\boldsymbol{X},D}\right| = \frac{|R_{Y\sim Z|\boldsymbol{X}}| - |R_{Y\sim D|\boldsymbol{X}} R_{D\sim Z|\boldsymbol{X}}|}{\sqrt{1 - R^2_{Y\sim D|\boldsymbol{X}}}\sqrt{1 - R^2_{D\sim Z|\boldsymbol{X}}}}. \tag{7.29}$$

Extending the previous arguments to multiple covariates is straightforward, since these results hold for any subset of $\boldsymbol{X}$.

7.1.3.2 Comparing the partial R^2 of covariates with the partial R^2 of the confounder

Now imagine the researcher is willing to make a more elaborate type of claim. For instance, the researcher believes that omitting X_j increases the mean squared error of the full treatment regression more than omitting Z. This means that, $R^2_{D\sim\boldsymbol{X}_{-j}+Z} < R^2_{D\sim\boldsymbol{X}}$, where $\boldsymbol{X}_{-j}$ represents all variables in $\boldsymbol{X}$ except X_j. If we now subtract of both sides $R^2_{D\sim\boldsymbol{X}_{-j}}$ and further divide them by $1 - R^2_{D\sim\boldsymbol{X}_{-j}}$, this gives us,

$$\frac{R^2_{D\sim\boldsymbol{X}_{-j}+Z} - R^2_{D\sim\boldsymbol{X}_{-j}}}{1 - R^2_{D\sim\boldsymbol{X}_{-j}}} < \frac{R^2_{D\sim\boldsymbol{X}} - R^2_{D\sim\boldsymbol{X}_{-j}}}{1 - R^2_{D\sim\boldsymbol{X}_{-j}}}. \tag{7.30}$$

Which means that

$$R^2_{D\sim Z|\boldsymbol{X}_{-j}} < R^2_{D\sim X_j|\boldsymbol{X}_{-j}}. \tag{7.31}$$

That is, we can compare the strength of Z to X_j by assessing their relative contribution to the partial R^2 of the treatment regression given the remaining covariates. Generalizing this notion define,

$$k_D := \frac{R^2_{D\sim Z|\boldsymbol{X}_{-j}}}{R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}. \tag{7.32}$$

Our goal now is to re-express $R^2_{D\sim Z|\boldsymbol{X}}$ in terms of k_D.

Bounding $R^2_{D\sim Z|\boldsymbol{X}}$

From Equation 7.32 we have that $|R_{D\sim Z|\boldsymbol{X}_{-j}}| = \sqrt{k_D}|R_{D\sim X_j|\boldsymbol{X}_{-j}}|$. Also, the assumption that $Z \perp \boldsymbol{X}$ implies $R_{Z\sim X_j|\boldsymbol{X}_{-j}} = 0$. Combining these two results, and using the standard recursive definition of partial correlations, gives us

$$\begin{aligned}
\left|R_{D\sim Z|\boldsymbol{X}}\right| &= \left|\frac{R_{D\sim Z|\boldsymbol{X}_{-j}} - R_{D\sim X_j|\boldsymbol{X}_{-j}} R_{Z\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1 - R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}\sqrt{1 - R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (7.33)\\
&= \left|\frac{R_{D\sim Z|\boldsymbol{X}_{-j}}}{\sqrt{1 - R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (7.34)\\
&= \frac{\sqrt{k_D}\left|R_{D\sim X_j|\boldsymbol{X}_{-j}}\right|}{\sqrt{1 - R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}} && (7.35)\\
&= \sqrt{k_D}\left|f_{D\sim X_j|\boldsymbol{X}-j}\right|. && (7.36)
\end{aligned}$$

Hence,

$$R^2_{D\sim Z|\boldsymbol{X}} = k_D \times f^2_{D\sim X_j|\boldsymbol{X}-j}. \tag{7.37}$$

Also, notice that, since $R^2_{D\sim Z|\boldsymbol{X}} \leq 1$ this, means k_D cannot vary freely but rather is bounded by

$$k_D \leq \frac{1}{f^2_{D\sim X_j|\boldsymbol{X}-j}}. \tag{7.38}$$

As an example, if a researcher has a covariate that currently explains 50% of the residual variance of the treatment assignment (implying $f^2_{D\sim X_j|\boldsymbol{X}-j} = 1$), Equation 7.38 reveals it is *impossible* to have an *orthogonal* unobserved confounder Z stronger than that covariate.

Using multiple covariates. Now let us generalize the previous bound to multiple covariates. Let this set of covariates be $\boldsymbol{X}_{(1\dots j)} = \{X_1, \dots, X_j\}$. We will denote the complement of this set $\boldsymbol{X}_{-(1\dots j)}$. Thus, k_D now is defined as

$$k_D := \frac{R^2_{D\sim Z|\boldsymbol{X}_{-(1\dots j)}}}{R^2_{D\sim \boldsymbol{X}_{(1\dots j)}|\boldsymbol{X}_{-(1\dots j)}}}. \tag{7.39}$$

Applying the recursive definition of partial correlation to, $R_{D\sim Z|\boldsymbol{X}}$, $R_{D\sim Z|\boldsymbol{X}_{-(1)}}$, $R_{D\sim Z|\boldsymbol{X}_{-(1,2)}}$, up to $R_{D\sim Z|\boldsymbol{X}_{-(1,\dots,j)}}$, and recalling the orthogonality of Z with $\boldsymbol{X}$, we have that,

$$R_{D\sim Z|\boldsymbol{X}} = \frac{R_{D\sim Z|\boldsymbol{X}_{-(1,\dots,j)}}}{\sqrt{1-R^2_{D\sim X_1|\boldsymbol{X}_{-(1)}}}\sqrt{1-R^2_{D\sim X_2|\boldsymbol{X}_{-(1,2)}}}\cdots\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-(1,\dots,j)}}}}. \tag{7.40}$$

Since, $R^2_{D\sim Z|\boldsymbol{X}_{-(1\dots j)}} = k_D R^2_{D\sim \boldsymbol{X}_{(1\dots j)}|\boldsymbol{X}_{-(1\dots j)}}$, we obtain,

$$\left|R_{D\sim Z|\boldsymbol{X}}\right| = \frac{\sqrt{k_D}\left|R_{D\sim \boldsymbol{X}_{(1\dots j)}|\boldsymbol{X}_{-(1\dots j)}}\right|}{\sqrt{1-R^2_{D\sim X_1|\boldsymbol{X}_{-(1)}}}\sqrt{1-R^2_{D\sim X_2|\boldsymbol{X}_{-(1,2)}}}\cdots\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-(1,\dots,j)}}}}. \tag{7.41}$$

We can simplify this further by noticing the denominator is simply $\sqrt{1-R^2_{D\sim \boldsymbol{X}_{(1\dots j)}|\boldsymbol{X}_{-(1\dots j)}}}$

$$\left|R_{D\sim Z|\boldsymbol{X}}\right| = \frac{\sqrt{k_D}\left|R_{D\sim \boldsymbol{X}_{(1\dots j)}|\boldsymbol{X}_{-(1\dots j)}}\right|}{\sqrt{1-R^2_{D\sim \boldsymbol{X}_{(1\dots j)}|\boldsymbol{X}_{-(1\dots j)}}}} = \sqrt{k_D}\left|f_{D\sim \boldsymbol{X}_{(1\dots j)}|\boldsymbol{X}_{-(1\dots j)}}\right|. \tag{7.42}$$

Bounding $R_{Y\sim Z|D,\boldsymbol{X}}$

We have two ways of bounding $R_{Y\sim Z|D,\boldsymbol{X}}$, making comparisons conditional or not conditional on D.

Comparisons not conditioning on D. As in the previous derivation, define,

$$k_Y := \frac{R^2_{Y\sim Z|\boldsymbol{X}_{-(1\dots j)}}}{R^2_{Y\sim \boldsymbol{X}_{(1\dots j)}|\boldsymbol{X}_{-(1\dots j)}}}. \tag{7.43}$$

That is, we are asking the researcher to compare the explanatory power of the confounder against the explanatory power of $\boldsymbol{X}_{(1\dots j)}$ with respect to the outcome, conditioning on the remaining covariates $\boldsymbol{X}_{-(1\dots j)}$ but *not conditioning* on the treatment. Using the same recursive argument as before, we obtain

$$\left|R_{Y\sim Z|\boldsymbol{X}}\right| = \sqrt{k_Y}\left|f_{Y\sim \boldsymbol{X}_{(1\dots j)}|\boldsymbol{X}_{-(1\dots j)}}\right|. \tag{7.44}$$

We can now bound $R^2_{Y\sim Z|D,\boldsymbol{X}}$ by noting again that

$$R_{Y\sim Z|D,\boldsymbol{X}} = \frac{R_{Y\sim Z|\boldsymbol{X}} - R_{Y\sim D|\boldsymbol{X}}R_{D\sim Z|\boldsymbol{X}}}{\sqrt{1-R^2_{Y\sim D|\boldsymbol{X}}}\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}}}}, \tag{7.45}$$

then using the same argument as in 7.1.3.2.

Comparisons conditioning on D. Here we have that k_D is defined as before, but k_Y compares the explanatory power of the confounder against the explanatory power of a covariate X_j with respect to the outcome, conditioning on both the remaining covariates $\boldsymbol{X}_{-(1\dots j)}$ and the treatment, that is,

$$k_D := \frac{R^2_{D\sim Z|\boldsymbol{X}_{-j}}}{R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}, \qquad k_Y := \frac{R^2_{Y\sim Z|\boldsymbol{X}_{-j},D}}{R^2_{Y\sim X_j|\boldsymbol{X}_{-j},D}}. \tag{7.46}$$

To bound $R^2_{Y\sim Z|D,\boldsymbol{X}}$, we first need to investigate $R_{Z\sim X_j|\boldsymbol{X}_{-j},D}$. Expanding the partial correlation gives us

$$\begin{aligned}
\left|R_{Z\sim X_j|\boldsymbol{X}_{-j},D}\right| &= \left|\frac{R_{Z\sim X_j|\boldsymbol{X}_{-j}} - R_{D\sim Z|\boldsymbol{X}_{-j}}R_{D\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}_{-j}}}\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (7.47)\\
&= \left|\frac{R_{D\sim Z|\boldsymbol{X}_{-j}}R_{D\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}_{-j}}}\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (7.48)\\
&= \left|\frac{\sqrt{k_D}R_{D\sim X_j|\boldsymbol{X}_{-j}}R_{D\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1-k_D R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (7.49)\\
&= \left|f_{K_D} \times f_{D\sim X_j|\boldsymbol{X}-j}\right|. && (7.50)
\end{aligned}$$

where, f_{K_D} is defined to be,

$$f_{K_D} := \frac{\sqrt{k_D}R_{D\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1-k_D R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}. \tag{7.51}$$

Combining theses results and Equation 7.46 we can proceed to bound $R_{Y\sim Z|D,\boldsymbol{X}}$:

$$
\begin{aligned}
|R_{Y\sim Z|D,\boldsymbol{X}}| &= \left|\frac{R_{Y\sim Z|\boldsymbol{X}_{-j},D} - R_{Y\sim X_j|\boldsymbol{X}_{-j}D} R_{Z\sim X_j|\boldsymbol{X}_{-j},D}}{\sqrt{1-R^2_{Y\sim X_j|\boldsymbol{X}_{-j}D}}\sqrt{1-R^2_{Z\sim X_j|\boldsymbol{X}_{-j},D}}}\right| && (7.52)\\
&\leq \frac{\left|R_{Y\sim Z|\boldsymbol{X}_{-j},D}\right| + \left|R_{Y\sim X_j|\boldsymbol{X}_{-j}D}\right|\left|R_{Z\sim X_j|\boldsymbol{X}_{-j},D}\right|}{\sqrt{1-R^2_{Y\sim X_j|\boldsymbol{X}_{-j}D}}\sqrt{1-R^2_{Z\sim X_j|\boldsymbol{X}_{-j},D}}} && (7.53)\\
&= \frac{\sqrt{k_Y}\left|R_{Y\sim X_j|\boldsymbol{X}_{-j}D}\right| + \left|R_{Y\sim X_j|\boldsymbol{X}_{-j}D}\right|\left|f_{K_D} \times f_{D\sim X_j|\boldsymbol{X}-j}\right|}{\sqrt{1-R^2_{Y\sim X_j|\boldsymbol{X}_{-j}D}}\sqrt{1-f^2_{K_D} \times f^2_{D\sim X_j|\boldsymbol{X}-j}}} && (7.54)\\
&= \left(\frac{\sqrt{k_Y} + \left|f_{K_D} \times f_{D\sim X_j|\boldsymbol{X}-j}\right|}{\sqrt{1-f^2_{K_D} \times f^2_{D\sim X_j|\boldsymbol{X}-j}}}\right)\left(\frac{\left|R_{Y\sim X_j|\boldsymbol{X}_{-j}D}\right|}{\sqrt{1-R^2_{Y\sim X_j|\boldsymbol{X}_{-j}D}}}\right) && (7.55)\\
&= \eta\left|f_{Y\sim X_j|\boldsymbol{X}_{-j},D}\right|. && (7.56)
\end{aligned}
$$

Hence, we have that,

$$R^2_{Y\sim Z|D,\boldsymbol{X}} \leq \eta^2 f^2_{Y\sim X_j|\boldsymbol{X}_{-j},D}, \tag{7.57}$$

where $\eta = \frac{\sqrt{k_Y}+\left|f_{K_D}\times f_{D\sim X_j|\boldsymbol{X}-j}\right|}{\sqrt{1-f^2_{K_D}\times f^2_{D\sim X_j|\boldsymbol{X}-j}}}$. Note the bound is tight. Without further assumptions, we can create an unobserved confounder Z that makes the inequality step in 7.53 an equality. One can extend this to multiple covariates by iteratively applying the recursive definition of partial correlation.

7.1.4 Some numerical examples of informal benchmarking

Here we show how the informal benchmarking practices proposed in [58, 59] and in [26] could lead users to erroneous conclusions. Starting with [26], consider the simulation in the `R` code presented in the left hand side of Figure 7.1. Note the unobserved confounder Z is exactly like X in terms of its association with the treatment D and the outcome Y; moreover, we also have that $Z \perp X$. Finally, note that, by construction, the unobserved confounder Z (which is as strong as X) is sufficient to bring the effect estimate down to zero. The right hand side

```
# cleans workspace
rm(list = ls())

# set seed for reproducibility
set.seed(10)

# loads packages
library(treatSens)
library(konfound)

# simulates data
n <- 500
x <- rnorm(n)
z <- rnorm(n)
d <- x + z + rnorm(n)
y <- x + z + rnorm(n)

# Carnegie et al method
sense <- treatSens(y ~ d + x)
sensPlot(sense)

# Frank's method
model <- lm(y ~ d + x)

## computes impact threshold
konfound(model, tested_variable = "d",
                alpha = 0.05)

## "observed impact" of X
cor(x, d)*cor(x, y)
```

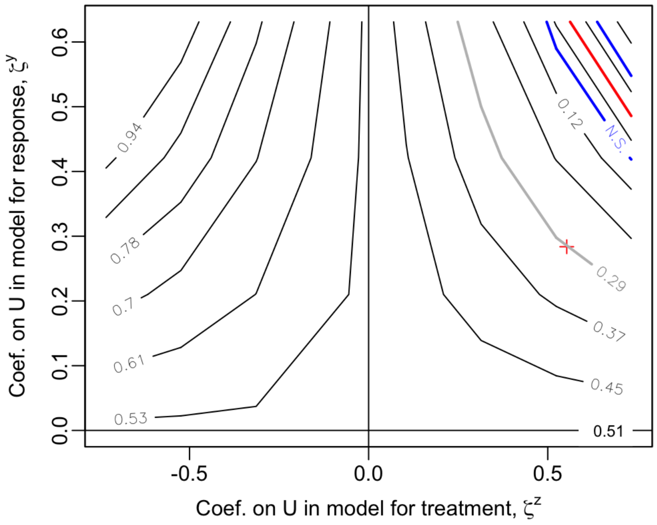


Figure 7.1: Examples of informal benchmarking.

Note: Code (left) and plot (right) for the incorrect informal benchmark bound produced from the methods of Carnegie, Harada and Hill (2006). Note the informal benchmark would lead one to *incorrectly* conclude that an unobserved confounder Z exactly like X would not be sufficient to explain away the estimated effect, when in fact it would (as shown in the red "x" mark). Code for [58] and [59] is also shown in the left.

of Figure 7.1 shows the output of [26] companion software, the R package treatSens [27].[1] Note it incorrectly claims that the effect estimate would be robust to a confounder as strong as X (benchmark shown in the red "x" mark).

Now moving to [58] and [59], one would first compute the "impact threshold" of a confounding variable and compare this to the "observed" *impact* of the covariate X. In the same simulation of Figure 7.1, these calculation are shown in the last part of the code (using the R package konfound). One then obtains an impact threshold of 0.469 (considering statistical significance of 5%), which, when contrasted with the "observed impact" of X, $R_{Y\sim X} \times R_{D\sim X} = 0.314$, would lead an investigator to erroneously conclude that an unobserved confounder as strong as X would not be sufficient to explain away the estimate.

[1] As of 14 October 2019, the R package was removed from CRAN for lack of maitainance; archived versions can still be found in https://cran.r-project.org/web/packages/treatSens/index.html.

7.2 Appendix for Chapter 3

7.2.1 Main estimators for IV

For ease of reference, in this section we show in more detail some of the algebraic identities (and differences) of the main approaches to IV estimation.

7.2.1.1 Indirect Least Squares (ILS)

Point Estimate. The ILS estimate is defined as the ratio of the reduced-form and first-stage estimates

$$\hat{\tau}_{\mathrm{ILS}} := \frac{\hat{\lambda}}{\hat{\theta}} \tag{7.58}$$

Inference. Inference in the ILS framework is usually performed using the delta-method, with estimated variance

$$\widehat{\mathrm{var}}(\hat{\tau}_{\mathrm{ILS}}) := \frac{1}{\hat{\theta}^2}\left(\widehat{\mathrm{var}}(\hat{\lambda}) + \hat{\tau}^2\widehat{\mathrm{var}}(\hat{\theta}) - 2\hat{\tau}\widehat{\mathrm{cov}}(\hat{\lambda},\hat{\theta})\right) \tag{7.59}$$

where, using the FWL formulation,

$$\widehat{\mathrm{var}}(\hat{\lambda}) = \frac{\mathrm{var}(Y^{\perp Z,\boldsymbol{X},\boldsymbol{W}})}{\mathrm{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} \times \mathrm{df}^{-1}, \qquad \widehat{\mathrm{var}}(\hat{\theta}) = \frac{\mathrm{var}(D^{\perp Z,\boldsymbol{X},\boldsymbol{W}})}{\mathrm{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} \times \mathrm{df}^{-1} \tag{7.60}$$

are the estimated variances of the reduced form and first stage, and

$$\widehat{\mathrm{cov}}(\hat{\lambda},\hat{\theta}) = \frac{\mathrm{cov}(Y^{\perp Z,\boldsymbol{X},\boldsymbol{W}}, D^{\perp Z,\boldsymbol{X},\boldsymbol{W}})}{\mathrm{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} \times \mathrm{df}^{-1} \tag{7.61}$$

is the estimated covariance of $\hat{\lambda}$ and $\hat{\theta}$.

7.2.1.2 Two-Stage Least Squares (2SLS)

Point Estimate. By the FWL theorem, the 2SLS point estimate can be written as

$$\hat{\tau}_{\text{2SLS}} := \frac{\text{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}}, \widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})}{\text{var}(\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})} \tag{7.62}$$

In the just-identified case, the ILS and 2SLS point estimates are numerically identical. Expanding $\widehat{D}$ we have that

$$\begin{aligned} \hat{\tau}_{\text{2SLS}} &= \frac{\text{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}}, \widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})}{\text{var}(\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})} = \frac{\text{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}}, \hat{\theta} Z^{\perp \boldsymbol{X},\boldsymbol{W}})}{\text{var}(\hat{\theta} Z^{\perp \boldsymbol{X},\boldsymbol{W}})} \qquad (7.63) \\ &= \frac{\hat{\theta} \times \text{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}}, Z^{\perp \boldsymbol{X},\boldsymbol{W}})}{\hat{\theta}^2 \times \text{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} = \frac{\hat{\lambda}}{\hat{\theta}} \qquad (7.64) \end{aligned}$$

Which establishes the equality $\hat{\tau}_{\text{2SLS}} = \hat{\tau}_{\text{ILS}}$.

Inference. By the FWL theorem, the standard two-stage least squares estimate of the variance can be written as

$$\widehat{\text{var}}(\hat{\tau}_{\text{2SLS}}) := \frac{\text{var}(Y^{\perp \boldsymbol{X},\boldsymbol{W}} - \hat{\tau} D^{\perp \boldsymbol{X},\boldsymbol{W}})}{\text{var}(\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})} \times \text{df}^{-1} \tag{7.65}$$

As with the point estimate, for the just-identified case, the estimated variance of ILS and 2SLS are numerically identical. To see why, note the denominator of Equation 7.65 can be expanded to

$$\text{var}(\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}}) = \text{var}(\hat{\theta} Z^{\perp \boldsymbol{X},\boldsymbol{W}}) = \hat{\theta}^2 \text{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}}) \tag{7.66}$$

Finally, the numerator can be written as,

$$\begin{aligned}
\mathrm{var}(Y^{\perp \boldsymbol{X},\boldsymbol{W}} - \hat{\tau} D^{\perp \boldsymbol{X},\boldsymbol{W}}) &= \mathrm{var}(Y^{\perp \boldsymbol{X},\boldsymbol{W}} - \hat{\tau}(\hat{\theta} Z^{\boldsymbol{X},\boldsymbol{W}} + D^{\perp Z,\boldsymbol{X},\boldsymbol{W}})) \qquad (7.67)\\
&= \mathrm{var}((Y^{\perp \boldsymbol{X},\boldsymbol{W}} - \hat{\lambda} Z^{\boldsymbol{X},\boldsymbol{W}}) - \hat{\tau} D^{\perp Z,\boldsymbol{X},\boldsymbol{W}}) \qquad (7.68)\\
&= \mathrm{var}(Y^{\perp Z,\boldsymbol{X},\boldsymbol{W}} - \hat{\tau} D^{\perp Z,\boldsymbol{X},\boldsymbol{W}}) \qquad (7.69)\\
&= \mathrm{var}(Y^{\perp Z,\boldsymbol{X},\boldsymbol{W}}) + \hat{\tau}^2 \mathrm{var}(D^{\perp Z,\boldsymbol{X},\boldsymbol{W}}) - 2\hat{\tau}\mathrm{cov}(Y^{\perp Z,\boldsymbol{X},\boldsymbol{W}}, D^{\perp Z,\boldsymbol{X},\boldsymbol{W}}) \qquad (7.70)
\end{aligned}$$

Plugging in Equations 7.70 and 7.66 back in Equation 7.65, then using Equations 7.60 and 7.61 establishes the desired equality.

7.2.1.3 Anderson-Rubin (AR)

Point Estimate. We define the Anderson-Rubin point estimate to be the value of τ_0 that makes $\hat{\phi} = 0$, ie,

$$\hat{\tau}_{\mathrm{AR}} = \{\tau_0;\ \hat{\phi}_{\tau_0} = 0\} \tag{7.71}$$

Resorting again to the FWL theorem, we can write the regression coefficient of the AR regression, $\hat{\phi}_{\tau_0}$, as a function of the regression coefficients of the first stage and reduced form,

$$\begin{aligned}
\hat{\phi}_{\tau_0} &= \frac{\mathrm{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}} - \tau_0 D^{\perp \boldsymbol{X},\boldsymbol{W}}, Z^{\perp \boldsymbol{X},\boldsymbol{W}})}{\mathrm{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} \qquad (7.72)\\
&= \frac{\mathrm{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}}, Z^{\perp \boldsymbol{X},\boldsymbol{W}})}{\mathrm{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} - \tau_0 \frac{\mathrm{cov}(D^{\perp \boldsymbol{X},\boldsymbol{W}}, Z^{\perp \boldsymbol{X},\boldsymbol{W}})}{\mathrm{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} \qquad (7.73)\\
&= \hat{\lambda} - \tau_0 \hat{\theta} \qquad (7.74)
\end{aligned}$$

Thus solving for the condition $\hat{\phi}_{\tau_0} = 0$ gives us

$$\hat{\tau}_{AR} = \frac{\hat{\lambda}}{\hat{\theta}} \tag{7.75}$$

Which establishes the equality $\hat{\tau}_{AR} = \hat{\tau}_{ILS}$. Therefore, all the point estimates of ILS, 2SLS and AR are numerically identical.

Inference. The AR confidence interval with significance level α is defined as all values of τ_0 such that we cannot reject the null hypothesis $H_0 : \phi_{\tau_0} = 0$ at the chosen significance level

$$\text{CI}_{1-\alpha}(\tau) = \{\tau_0; t^2_{\hat{\phi}_{\tau_0}} \leq t^{*2}_{\alpha,\text{df}}\} \tag{7.76}$$

This confidence interval can be obtained analytically as functions of the estimates of the first-stage and reduced form regressions. As shown in Equation 7.74, $\hat{\phi}_{\tau_0}$ can be written as the linear combination

$$\hat{\phi}_{\tau_0} = \hat{\lambda} - \tau_0\hat{\theta} \tag{7.77}$$

Likewise, by the FWL theorem, the estimated variance is given by

$$\widehat{\text{var}}(\hat{\phi}_{\tau_0}) = \frac{\text{var}(Y^{\perp Z,\boldsymbol{X},\boldsymbol{W}} - \tau_0 D^{\perp Z,\boldsymbol{X},\boldsymbol{W}})}{\text{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} \times \text{df}^{-1} \tag{7.78}$$

$$= \left(\frac{\text{var}(Y^{\perp Z,\boldsymbol{X},\boldsymbol{W}})}{\text{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} + \tau_0^2\frac{\text{var}(D^{\perp Z,\boldsymbol{X},\boldsymbol{W}})}{\text{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} - 2\tau_0\frac{\text{cov}(Y^{\perp Z,\boldsymbol{X},\boldsymbol{W}}, D^{\perp Z,\boldsymbol{X},\boldsymbol{W}})}{\text{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})}\right) \times \text{df}^{-1} \tag{7.79}$$

$$= \widehat{\text{var}}(\hat{\lambda}) + \tau_0^2\widehat{\text{var}}(\hat{\theta}) - 2\tau_0\widehat{\text{cov}}(\hat{\lambda}, \hat{\theta}) \tag{7.80}$$

Thus, we obtain that the t-value $t_{\hat{\phi}_{\tau_0}}$ for testing the null hypothesis $H_0 : \phi_{\tau_0} = 0$ equals to

$$t_{\hat{\phi}_{\tau_0}} = \frac{\hat{\lambda} - \tau_0\hat{\theta}}{\sqrt{\widehat{\text{var}}(\hat{\lambda}) + \tau_0^2\widehat{\text{var}}(\hat{\theta}) - 2\tau_0\widehat{\text{cov}}(\hat{\lambda}, \hat{\theta})}} \tag{7.81}$$

And our task is to find all values of τ_0 such that the following inequality holds

$$\frac{(\hat{\lambda} - \tau_0\hat{\theta})^2}{\widehat{\text{var}}(\hat{\lambda}) + \tau_0^2\widehat{\text{var}}(\hat{\theta}) - 2\tau_0\widehat{\text{cov}}(\hat{\lambda}, \hat{\theta})} \leq t^{*2}_{\alpha,\text{df}} \tag{7.82}$$

First, note that the empty set is not possible here. If we pick $\tau_0 = \hat{\tau}_{\text{AR}}$, then the numerator in Equation 7.82 is zero, and the inequality trivially holds—therefore, the point-estimate is always included in the confidence interval. Now squaring and rearranging terms we obtain

$$\underbrace{\left(\hat{\theta}^2 - \widehat{\text{var}}(\hat{\theta}) \times t^{*2}_{\alpha,\text{df}}\right)}_{a} \tau_0^2 + 2 \underbrace{\left(\widehat{\text{cov}}(\hat{\lambda}, \hat{\theta}) \times t^{*2}_{\alpha,\text{df}} - \hat{\lambda}\hat{\theta}\right)}_{b} \tau_0 + \underbrace{\left(\hat{\lambda}^2 - \widehat{\text{var}}(\hat{\lambda}) \times t^{*2}_{\alpha,\text{df}}\right)}_{c} \leq 0 \quad (7.83)$$

Our task has simplified to find all values of τ_0 that makes the above quadratic equation, with coefficients a, b and c, non-positive. As discussed in Section 3.4.2.2, this confidence interval can take three different forms, depending on the instrument strength: (i) finite and connected, (ii) the union two disjoint half lines; or, (iii) the whole real line.

7.2.1.4 Fieller's theorem

Fieller's proposal to test the null hypothesis $H_0 : \tau = \tau_0$ is to construct the linear combination $\hat{\phi}_{\tau_0} = \hat{\lambda} - \tau_0\hat{\theta}$, and to test the null hypothesis $H_0 : \phi_{\tau_0} = 0$. The standard estimated variance for $\hat{\phi}_{\tau_0}$ equals Equation 7.80, resulting in a test statistic equal to Equation 7.81, and thus numerically identical to the AR approach.

7.2.2 OVB-adjusted critical values and set of compatible inferences

7.2.2.1 OVB-adjusted critical values

As in the main text, using the reduced form as an example, let $\text{LL}_{1-\alpha}(\lambda) := \hat{\lambda} - t^*_{\alpha,\text{df}-1} \times \widehat{\text{se}}(\hat{\lambda})$ be the lower limit of a $1-\alpha$ level confidence interval of the full reduced form regression, where $t^*_{\alpha,\text{df}-1}$ denotes the critical α-level threshold of the t-distribution with $\text{df} - 1$ degrees of freedom. Considering the direction of the bias that reduces the lower limit, Equations 3.24 and 3.26

imply

$$\begin{aligned}
\mathrm{LL}_{1-\alpha}(\lambda) &:= \hat{\lambda} - t^*_{\alpha,\mathrm{df}-1} \times \widehat{\mathrm{se}}(\hat{\lambda}) && (7.84)\\
&= \hat{\lambda}_{\mathrm{res}} - \mathrm{BF}\sqrt{\mathrm{df}} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) - t^*_{\alpha,\mathrm{df}-1} \times \mathrm{SEF}\sqrt{\mathrm{df}/(\mathrm{df}-1)} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) && (7.85)\\
&= \hat{\lambda}_{\mathrm{res}} - \left(\mathrm{SEF}\sqrt{\mathrm{df}/(\mathrm{df}-1)} \times t^*_{\alpha,\mathrm{df}-1} + \mathrm{BF}\sqrt{\mathrm{df}}\right) \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) && (7.86)
\end{aligned}$$

Similarly, now let $\mathrm{UL}_{1-\alpha}(\lambda)$ the upper limit of the confidence interval and consider the direction of the bias that increases the upper limit. By the same algebraic manipulations, we obtain

$$\mathrm{UL}_{1-\alpha}(\lambda) = \hat{\lambda}_{\mathrm{res}} + \left(\mathrm{SEF}\sqrt{\mathrm{df}/(\mathrm{df}-1)} \times t^*_{\alpha,\mathrm{df}-1} + \mathrm{BF}\sqrt{\mathrm{df}}\right) \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) \qquad (7.87)$$

Note that, in both Equations 7.86 and 7.87, the only part that depends on the omitted variable W is the common multiple of the observed standard error, which defines the new *OVB-adjusted critical value*,

$$t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} := \mathrm{SEF}\sqrt{\mathrm{df}/(\mathrm{df}-1)} \times t^*_{\alpha,\mathrm{df}-1} + \mathrm{BF}\sqrt{\mathrm{df}}. \qquad (7.88)$$

7.2.2.2 Compatible inferences given bounds on the partial R^2

Now suppose the analyst wishes to investigate the worst possible lower (or upper) limits of the confidence intervals induced by a confounder with strength no stronger than certain bounds, for instance, $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq R^{2\max}_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\max}_{Z\sim W|\boldsymbol{X}}$. As per the last section, this amounts to finding the largest *OVB-adjusted critical value* induced by an omitted variable W with at most such strength. That is, we need to solve the following maximization problem

$$\max_{R^2_{Y\sim W|Z,\boldsymbol{X}},R^2_{Z\sim W|\boldsymbol{X}}} t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \quad \text{s.t.} \quad R^2_{Y\sim W|Z,\boldsymbol{X}} \leq R^{2\max}_{Y\sim W|Z,\boldsymbol{X}}, \quad R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\max}_{Z\sim W|\boldsymbol{X}} \qquad (7.89)$$

Dividing $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ by $\sqrt{\mathrm{df}}$ and letting $f^{*}_{\alpha,\mathrm{df}-1} := t^{*}_{\alpha,\mathrm{df}-1}/\sqrt{\mathrm{df}-1}$, we see that the derivative of $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ with respect to $R^2_{Z\sim W|\boldsymbol{X}}$ is always increasing, since

$$\begin{aligned}
\frac{\partial(t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}/\sqrt{\mathrm{df}})}{\partial R^2_{Z\sim W|\boldsymbol{X}}} &= \frac{\partial\,\mathrm{BF}}{\partial R^2_{Z\sim W|\boldsymbol{X}}} + f^{*}_{\alpha,\mathrm{df}-1} \times \frac{\partial\,\mathrm{SEF}}{\partial R^2_{Z\sim W|\boldsymbol{X}}} \qquad (7.90)\\
&= \frac{(R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2}}{2(1-R^2_{Z\sim W|\boldsymbol{X}})^{3/2}(R^2_{Z\sim W|\boldsymbol{X}})^{1/2}} + f^{*}_{\alpha,\mathrm{df}-1}\frac{(1-R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2}}{2(1-R^2_{Z\sim W|\boldsymbol{X}})^{3/2}} \qquad (7.91)\\
&= \frac{(R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2} + f^{*}_{\alpha,\mathrm{df}-1}(1-R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2}(R^2_{Z\sim W|\boldsymbol{X}})^{1/2}}{2(1-R^2_{Z\sim W|\boldsymbol{X}})^{3/2}(R^2_{Z\sim W|\boldsymbol{X}})^{1/2}} \geq 0 \qquad (7.92)
\end{aligned}$$

Therefore, the "optimal" $R^{2*}_{Z\sim W|\boldsymbol{X}}$ (the one the minimizes (maximizes) the lower (upper) limit of the confidence interval) always reaches the bound. However, the same is not true for the derivative with respect to $R^2_{Y\sim W|Z,\boldsymbol{X}}$. To see that, write,

$$\begin{aligned}
\frac{\partial(t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}/\sqrt{\mathrm{df}})}{\partial R^2_{Y\sim W|Z,\boldsymbol{X}}} &= \frac{\partial\,\mathrm{BF}}{\partial R^2_{Y\sim W|Z,\boldsymbol{X}}} + f^{*}_{\alpha,\mathrm{df}-1} \times \frac{\partial\,\mathrm{SEF}}{\partial R^2_{Y\sim W|Z,\boldsymbol{X}}} \qquad (7.93)\\
&= \frac{(R^2_{Z\sim W|\boldsymbol{X}})^{1/2}}{2(1-R^2_{Z\sim W|\boldsymbol{X}})^{1/2}(R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2}} + \frac{-f^{*}_{\alpha,\mathrm{df}-1}}{2(1-R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2}(1-R^2_{Z\sim W|\boldsymbol{X}})^{1/2}} \\
&\qquad (7.94)\\
&= \frac{(R^2_{Z\sim W|\boldsymbol{X}})^{1/2}(1-R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2} - f^{*}_{\alpha,\mathrm{df}-1}(R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2}}{2(R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2}(1-R^2_{Y\sim W|Z,\boldsymbol{X}})^{1/2}(1-R^2_{Z\sim W|\boldsymbol{X}})^{1/2}} \qquad (7.95)
\end{aligned}$$

That is, due to the variance reduction factor of the omitted variable (VRF in Equation 3.26), it could be the case that increasing $R^2_{Y\sim W|Z,\boldsymbol{X}}$ reduces the standard error more than enough to compensate for the increase in bias, resulting in tighter confidence intervals.

We have, thus, two cases. First, consider the case in which the optimal point reaches both bounds. In that case, the numerator of Equation 7.95 must be positive when evaluated at the solution. Rearranging and squaring, we see that this happens when

$$R^{2\,\max}_{Z\sim W|\boldsymbol{X}} \geq f^{*2}_{\alpha,\mathrm{df}-1} \times f^{2\,\max}_{Y\sim W|Z,\boldsymbol{X}} \tag{7.96}$$

Clearly, when considering the sensitivity of the point estimate, we have $f^{*}_{\alpha,\mathrm{df}-1} = 0$, and the condition always holds. If condition of Equation 7.96 fails, then the optimal $R^{2*}_{Y\sim W|Z,\boldsymbol{X}}$ will be an interior point. This will happen when the numerator of Equation 7.95 equals zero. Since we know $R^2_{Z\sim W|\boldsymbol{X}}$ reaches its maximum, the optimal $R^{2*}_{Y\sim W|Z,\boldsymbol{X}}$ will be,

$$R^{2*}_{Y\sim W|Z,\boldsymbol{X}} = \frac{R^{2\,\max}_{Z\sim W|\boldsymbol{X}}}{f^{*2}_{\alpha,\mathrm{df}-1} + R^{2\,\max}_{Z\sim W|\boldsymbol{X}}} \tag{7.97}$$

Denoting the solution to the optimization problem as $t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$, the *most extreme possible* lower and upper limits after adjusting for W are given by

$$\mathrm{LL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda) = \hat{\lambda}_{\mathrm{res}} - t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}), \quad \mathrm{UL}^{\max}_{1-\alpha,\boldsymbol{R}^2} = \hat{\lambda}_{\mathrm{res}} + t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) \tag{7.98}$$

And interval composed of such limits

$$\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda) = \left[\mathrm{LL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda), \quad \mathrm{UL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)\right] \tag{7.99}$$

Defines the set of compatible inferences given the bounds on the partial R^2, $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq R^{2\,\max}_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\,\max}_{Z\sim W|\boldsymbol{X}}$.

7.2.3 (Extreme) Robustness Values

7.2.3.1 The Extreme Robustness Value

The *Extreme Robustness Value* $\mathrm{XRV}_{q^*,\alpha}(\lambda)$ is defined as the greatest lower bound XRV on the sensitivity parameter $R^2_{Z\sim W|\boldsymbol{X}}$, while keeping the parameter $R^2_{Y\sim W|Z,\boldsymbol{X}}$ unconstrained, such that the null hypothesis that a change of $(100 \times q)\%$ of the original estimate, $H_0 : \lambda = (1 - q^*)\hat{\lambda}_{\mathrm{res}}$, is not rejected at the α level:

$$\mathrm{XRV}_{q^*,\alpha}(\lambda) := \inf\left\{\mathrm{XRV};\ (1 - q^*)\hat{\lambda}_{\mathrm{res}} \in \mathrm{CI}^{\max}_{1-\alpha,1,\mathrm{XRV}}(\lambda)\right\} \tag{7.100}$$

First, consider the case where $f_{q^*}(\lambda) < f^*_{\alpha,\mathrm{df}-1}$. Note the XRV will be zero, since we already cannot reject the null hypothesis $H_0 : \lambda = (1-q^*)\hat{\lambda}_{\mathrm{res}}$ even assuming zero omitted variable bias. Next, note that, when $f^*_{\alpha,\mathrm{df}-1} > 0$, we can always pick a large enough value for $R^2_{Y\sim W|Z,\boldsymbol{X}}$ until condition 7.96 fails (since $f^2_{Y\sim W|Z,\boldsymbol{X}}$ is unbounded). Therefore, XRV will be given by an interior point solution. Using Equation 7.97 to express $t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ solely in terms of the optimal $R^2_{Z\sim W|\boldsymbol{X}}$, and solving for the value that gives us $(1-q^*)\hat{\lambda}_{\mathrm{res}}$, we obtain

$$\mathrm{XRV}_{q^*,\alpha}(\lambda) = \begin{cases} 0, & \text{if } \; f_{q^*}(\lambda) \leq f^*_{\alpha,\mathrm{df}-1} \\ \dfrac{f^2_{q^*}(\lambda) - f^{*2}_{\alpha,\mathrm{df}-1}}{1 + f^2_{q^*}(\lambda)}, & \text{otherwise.} \end{cases} \tag{7.101}$$

7.2.3.2 The Robustness Value

The *Robustness Value* $\mathrm{RV}_{q^*,\alpha}(\lambda)$ for not rejecting the null hypothesis that $H_0 : \lambda = (1-q^*)\hat{\lambda}_{\mathrm{res}}$, at the significance level α, is defined as

$$\mathrm{RV}_{q^*,\alpha}(\lambda) := \inf\left\{\mathrm{RV};\; (1-q^*)\hat{\lambda}_{\mathrm{res}} \in \mathrm{CI}^{\max}_{1-\alpha,\mathrm{RV},\mathrm{RV}}(\lambda)\right\} \tag{7.102}$$

Where now we consider both sensitivity parameters bounded by RV. Again, consider the case where $f_{q^*}(\lambda) < f^*_{\alpha,\mathrm{df}-1}$. The RV then must be zero, since we already cannot reject the null hypothesis $H_0 : \lambda = (1-q^*)\hat{\lambda}_{\mathrm{res}}$ given the current data. Next, let's consider the case when the bound on $R^2_{Y\sim W|Z,\boldsymbol{X}}$ is not biding—here our optimization problem reduces to the XRV case. Finally, we have the solution in which both coordinates achieve the bound, resulting in a quadratic equation as solved before for Chapter 2. We thus have,

$$\mathrm{RV}_{q^*,\alpha}(\lambda) = \begin{cases} 0, & \text{if } \; f_{q^*}(\lambda) \leq f^*_{\alpha,\mathrm{df}-1} \\ \dfrac{1}{2}\left(\sqrt{f^4_{q^*,\alpha}(\lambda) + 4f^2_{q^*,\alpha}(\lambda)} - f^2_{q^*,\alpha}(\lambda)\right), & \text{if } \; f^*_{\alpha,\mathrm{df}-1} < f_{q^*}(\lambda) < f^{*-1}_{\alpha,\mathrm{df}-1} \\ \mathrm{XRV}_{q^*,\alpha}(\lambda), & \text{otherwise.} \end{cases} \tag{7.103}$$

The condition $f_{q*}(\lambda) < f^{*-1}_{\alpha,\mathrm{df}-1}$, stems from the fact that the XRV solution cannot satisfy Equation 7.96. We now show that this is equivalent to the previous condition $\mathrm{RV}_{q^*,\alpha}(\lambda) > 1 - 1/f^2_{q^*}(\lambda)$. If $f_{q^*}(\lambda) < 1/f^*_{\alpha,\mathrm{df}-1}$ then,

$$\begin{aligned}
\mathrm{RV}_{q^*,\alpha}(\lambda) &= \frac{1}{2}\left(\sqrt{f^4_{q^*,\alpha}(\lambda) + 4f^2_{q^*,\alpha}(\lambda)} - f^2_{q^*,\alpha}(\lambda)\right) && (7.104)\\
&= \frac{1}{2}\left(\sqrt{(f_{q^*}(\lambda) - f^*_{\alpha,\mathrm{df}-1})^4 + 4(f_{q^*}(\lambda) - f^*_{\alpha,\mathrm{df}-1})^2} - (f_{q^*}(\lambda) - f^*_{\alpha,\mathrm{df}-1})^2\right) && (7.105)\\
&> \frac{1}{2}\left(\sqrt{(f_{q^*}(\lambda) - 1/f_{q^*}(\lambda))^4 + 4(f_{q^*}(\lambda) - 1/f_{q^*}(\lambda))^2} - (f_{q^*}(\lambda) - 1/f_{q^*}(\lambda))^2\right) && (7.106)\\
&= \frac{1}{2}\left(\sqrt{\left(\frac{f^2_q(\lambda)-1}{f_{q^*}(\lambda)}\right)^4 + 4\left(\frac{f^2_{q^*}(\lambda)-1}{f_{q^*}(\lambda)}\right)^2} - \left(\frac{f^2_{q^*}(\lambda)-1}{f_{q^*}(\lambda)}\right)^2\right) && (7.107)\\
&= \left(\frac{1}{2}\right)\left(\frac{f^2_{q^*}(\lambda)-1}{f^2_{q^*}(\lambda)}\right)\left(\sqrt{(f^2_q(\lambda)-1)^2 + 4f^2_{q^*}(\lambda)} - f^2_{q^*}(\lambda) + 1\right) && (7.108)\\
&= \left(\frac{1}{2}\right)\left(1 - 1/f^2_{q^*}(\lambda)\right)\left(\sqrt{f^4_q(\lambda) + 1 - 2f^2_{q^*}(\lambda) + 4f^2_{q^*}(\lambda)} - f^2_{q^*}(\lambda) + 1\right) && (7.109)\\
&= \left(\frac{1}{2}\right)\left(1 - 1/f^2_{q^*}(\lambda)\right)\left(\sqrt{f^4_q(\lambda) + 1 + 2f^2_{q^*}(\lambda)} - f^2_{q^*}(\lambda) + 1\right) && (7.110)\\
&= \left(\frac{1}{2}\right)\left(1 - 1/f^2_{q^*}(\lambda)\right)\left(f^2_{q^*}(\lambda) + 1 - f^2_{q^*}(\lambda) + 1\right) && (7.111)\\
&= 1 - 1/f^2_{q^*}(\lambda) && (7.112)
\end{aligned}$$

Therefore, $f_{q^*}(\lambda) < 1/f^*_{\alpha,\mathrm{df}-1} \implies \mathrm{RV}_{q^*,\alpha}(\lambda) > 1 - 1/f^2_{q^*}(\lambda)$. By the same argument one can derive $\mathrm{RV}_{q^*,\alpha}(\lambda) > 1 - 1/f^2_{q^*}(\lambda) \implies f_q(\lambda) > 1/f^*_{\alpha,\mathrm{df}-1}$. Hence, both conditions are equivalent. The new condition, however, is much simpler to verify.

7.2.4 Bounds on the strength of W

Let X_j be a specific covariate of the set $\boldsymbol{X}$. Now define

$$k_Z := \frac{R^2_{Z\sim W|\boldsymbol{X}_{-j}}}{R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}}, \qquad k_Y := \frac{R^2_{Y\sim W|Z,\boldsymbol{X}_{-j}}}{R^2_{Y\sim X_j|Z\boldsymbol{X}_{-j}}}. \tag{7.113}$$

Where $\boldsymbol{X}_{-j}$ is the set $\boldsymbol{X}$ excluding covariate X_j. Our goal in this section is to re-express (or bound) both sensitivity parameters as a function of the new parameters k_Z and k_Y and the observed data.

In Chapter 2 we showed how to obtains bounds for the strength of W under the assumption that $R^2_{W\sim X_j|\boldsymbol{X}_{-j}} = 0$, or, equivalently, when we consider the part of W not linearly explained by $\boldsymbol{X}$. This result may be particularly useful when considering both $\boldsymbol{X}$ and W as *causes* of Z, as in such cases contemplating the marginal orthogonality of W (or its part not explained by observed covariates) is more natural. Here we additionally provide bounds under the assumption that $R^2_{W\sim X_j|Z,\boldsymbol{X}_{-j}} = 0$. This condition may be helpful when contemplating the strength of W against X_j whenever these variables are *side-effects* of Z, instead of causes of Z. In such cases, reasoning about the marginal orthogonality of W with respect to $\boldsymbol{X}$ may not be natural, as Z itself is also a source of dependence between these variables.

We can thus start by re-expressing $R^2_{Y\sim W|Z,\boldsymbol{X}}$ in terms of k_Y, which in this case is straightforward. Using the recursive definition of partial correlations, and considering our two conditions $R^2_{W\sim X_j|Z,\boldsymbol{X}_{-j}} = 0$ and $R^2_{Y\sim W|Z,\boldsymbol{X}_{-j}} = k_Y R^2_{Y\sim X_j|Z\boldsymbol{X}_{-j}}$, we obtain

$$\left|R_{Y\sim W|Z,\boldsymbol{X}}\right| = \left|\frac{R_{Y\sim W|Z,\boldsymbol{X}_{-j}} - R_{Y\sim X_j|Z,\boldsymbol{X}_{-j}} R_{W\sim X_j|Z,\boldsymbol{X}_{-j}}}{\sqrt{1-R^2_{Y\sim X_j|Z,\boldsymbol{X}_{-j}}}\sqrt{1-R^2_{W\sim X_j|Z,\boldsymbol{X}_{-j}}}}\right| \tag{7.114}$$

$$= \left|\frac{R_{Y\sim W|Z,\boldsymbol{X}_{-j}}}{\sqrt{1-R^2_{Y\sim X_j|Z,\boldsymbol{X}_{-j}}}}\right| \tag{7.115}$$

$$= \left|\frac{\sqrt{k_Y} R_{Y\sim X_j|Z,\boldsymbol{X}_{-j}}}{\sqrt{1-R^2_{Y\sim X_j|Z,\boldsymbol{X}_{-j}}}}\right| \tag{7.116}$$

$$= \sqrt{k_Y}\left|f_{Y\sim X_j|Z,\boldsymbol{X}_{-j}}\right| \tag{7.117}$$

Hence,

$$R^2_{Y\sim W|Z,\boldsymbol{X}} = k_Y \times f^2_{Y\sim X_j|Z,\boldsymbol{X}_{-j}} \tag{7.118}$$

Moving to bound $R^2_{Z\sim W|\boldsymbol{X}}$, it is useful to first note that the conditions $R^2_{W\sim X_j|Z,\boldsymbol{X}_{-j}}=0$ and $R^2_{Z\sim W|\boldsymbol{X}_{-j}}=k_Z R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}$ allow us to re-express $R_{W\sim X_j|\boldsymbol{X}_{-j}}$ as a function of k_Z

$$R_{W\sim X_j|Z,\boldsymbol{X}_{-j}}=0 \implies \frac{R_{W\sim X_j|\boldsymbol{X}_{-j}}-R_{W\sim Z|\boldsymbol{X}_{-j}}R_{X_j\sim Z|\boldsymbol{X}_{-j}}}{\sqrt{1-R^2_{W\sim Z|\boldsymbol{X}_{-j}}}\sqrt{1-R^2_{X_j\sim Z|\boldsymbol{X}_{-j}}}}=0 \tag{7.119}$$

$$\implies R_{W\sim X_j|\boldsymbol{X}_{-j}}-R_{W\sim Z|\boldsymbol{X}_{-j}}R_{X_j\sim Z|\boldsymbol{X}_{-j}}=0 \tag{7.120}$$

$$\implies R_{W\sim X_j|\boldsymbol{X}_{-j}}=R_{W\sim Z|\boldsymbol{X}_{-j}}R_{X_j\sim Z|\boldsymbol{X}_{-j}} \tag{7.121}$$

$$\implies R_{W\sim X_j|\boldsymbol{X}_{-j}}=R_{Z\sim W|\boldsymbol{X}_{-j}}R_{Z\sim X_j|\boldsymbol{X}_{-j}} \tag{7.122}$$

$$\implies |R_{W\sim X_j|\boldsymbol{X}_{-j}}|=\sqrt{k_Z}R^2_{Z\sim X_j|\boldsymbol{X}_{-j}} \tag{7.123}$$

Now we can re-write $R^2_{Z\sim W|\boldsymbol{X}}$ using the recursive definition of partial correlations

$$\left|R_{Z\sim W|\boldsymbol{X}}\right|=\left|\frac{R_{Z\sim W|\boldsymbol{X}_{-j}}-R_{Z\sim X_j|\boldsymbol{X}_{-j}}R_{W\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1-R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}}\sqrt{1-R^2_{W\sim X_j|\boldsymbol{X}_{-j}}}}\right| \tag{7.124}$$

$$\le \frac{\left|R_{Z\sim W|\boldsymbol{X}_{-j}}\right|+\left|R_{Z\sim X_j|\boldsymbol{X}_{-j}}R_{W\sim X_j|\boldsymbol{X}_{-j}}\right|}{\sqrt{1-R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}}\sqrt{1-R^2_{W\sim X_j|\boldsymbol{X}_{-j}}}} \tag{7.125}$$

$$=\frac{\left|\sqrt{k_Z}R_{Z\sim X_j|\boldsymbol{X}_{-j}}\right|+\left|\sqrt{k_Z}R^3_{Z\sim X_j|\boldsymbol{X}_{-j}}\right|}{\sqrt{1-R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}}\sqrt{1-k_Z R^4_{Z\sim X_j|\boldsymbol{X}_{-j}}}} \tag{7.126}$$

$$=\left(\frac{\sqrt{k_Z}+\left|R^3_{Z\sim X_j|\boldsymbol{X}_{-j}}\right|}{\sqrt{1-k_Z R^4_{Z\sim X_j|\boldsymbol{X}_{-j}}}}\right)\times\left(\frac{\left|R_{Z\sim X_j|\boldsymbol{X}_{-j}}\right|}{\sqrt{1-R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}}}\right) \tag{7.127}$$

$$=\eta'|f_{Z\sim X_j|\boldsymbol{X}_{-j}}| \tag{7.128}$$

Hence we have that

$$R^2_{Z\sim W|\boldsymbol{X}}\le {\eta'}^2 f^2_{Z\sim X_j|\boldsymbol{X}_{-j}} \tag{7.129}$$

Where $\eta'=\left(\frac{\sqrt{k_Z}+\left|R^3_{Z\sim X_j|\boldsymbol{X}_{-j}}\right|}{\sqrt{1-k_Z R^4_{Z\sim X_j|\boldsymbol{X}_{-j}}}}\right)$.

7.3 Appendix for Chapter 4

7.3.1 Proof of propositions 1 and 2

The propositions follow directly from the definitions, but we state the proofs here for completeness. For proposition 1, first note ES is a functional of the covariance matrix Σ and it is by definition identifiable. Thus, if Q is identifiable, we can also uniquely compute Q from Σ and, since $B = ES - Q$, and each of its components is identifiable, B can also be uniquely computed from Σ and it is thus identifiable. Conversely, if B is identifiable, just write $Q = ES + B$, which means Q can be uniquely determined from Σ and it is also identifiable.

Proposition 2 follows the same argument. First note that if Q is θ-identifiable then we can write $B(\theta) = ES - Q(\theta)$ which is uniquely determined by Σ and θ, giving us a bias function parameterized in terms of θ. Conversely, if there exists a function $B(\theta)$ which, by definition, gives us a unique bias in terms of θ (and the data Σ), we can write $Q(\theta) = ES + B(\theta)$. This implies Q can be uniquely determined from Σ and θ and it is thus θ-identifiable.

7.3.2 Proof and pseudocode for Theorem 1

Theorem 1 (PushForward). *Given a linear SCM with graph G, covariance matrix Σ, a set of known directed edges $\mathcal{D}$, and known bidirected edge ε_{xy}, let the pair $\langle G', \Sigma' \rangle$ be constructed from G and Σ as follows:*

1. *$x \leftrightarrow y$ is removed and $\sigma'_{xy} = \sigma_{xy} - \varepsilon_{xy}$;*

2. *$\forall c \in Ch(x), c \neq y$, the bidirected edges $c \leftrightarrow y$ are added if they do not exist, and $\varepsilon'_{cy} = \varepsilon_{cy} + \lambda_{xc}\varepsilon_{xy}$;*

3. *$\forall z \in De(y)$, $z \neq x$, if there is an edge on any directed path from y to z that is not in $\mathcal{D}$, then z is removed from G'. For the remaining z, $\sigma'_{xz} = \sigma_{xz} - \varepsilon_{xy}\delta_{yz}$, where δ_{yz} is the sum of all directed paths from y to z;*

4. All other parameters and covariances remain the same.

Then if λ_{ab} is identifiable in G' it is $(\varepsilon_{xy}, \mathcal{D})$-identifiable in G.

Before moving forward, we use a couple definitions from the literature, which make reasoning about paths in the graph easier:

Definition 3. *[57] A path π from v to w is a* ***trek*** *if it has no colliding arrowheads, that is, π is of the form:*

$$v \leftarrow \ldots \leftarrow \leftrightarrow \rightarrow \ldots \rightarrow w$$

$$v \leftarrow \ldots \leftarrow k \rightarrow \ldots \rightarrow w$$

$$v \leftarrow \ldots \leftarrow w$$

$$v \rightarrow \ldots \rightarrow w$$

Definition 4. *[57] A* ***trek monomial*** *$\pi(\Lambda, \mathcal{E})$ for trek π is defined as the product of the structural parameters along the trek, multiplied by the trek's top error term covariance.*

In particular, if π does not contain a bidirected edge[2],

$$\pi(\Lambda, \mathcal{E}) = \varepsilon_k^2 \prod_{x \rightarrow y \in \pi} \lambda_{xy}$$

where k is the node at the "top" of the trek (it has no incoming edges). If the trek contains bidirected edge ε_{ab}, then

$$\pi(\Lambda, \mathcal{E}) = \varepsilon_{ab} \prod_{x \rightarrow y \in \pi} \lambda_{xy}$$

Lemma 1. *[57] The covariance between v and w, σ_{vw} can be written as the sum of the trek monomials of all treks between v and w ($\mathcal{T}_{vw}$):*

[2] Note also that we can have a trek from v to v, including a trek that takes no edges at all, which would be simply ε_v^2

Algorithm 2 PF - PushForward

```
function PF($G, \Sigma, \mathcal{D}, \varepsilon_{xy}, x$)
    initialize $\langle G', \Sigma' \rangle \leftarrow \langle G, \Sigma \rangle$
    update $\varepsilon'_{xy} \leftarrow 0$ in $G'$ and $\sigma'_{xy} \leftarrow \sigma'_{xy} - \varepsilon_{xy}$ in $\Sigma'$
    for each $c \in Ch(x)$ do
        update $\varepsilon'_{cy} \leftarrow \varepsilon'_{cy} + \lambda_{xc}\varepsilon_{xy}$ in $G'$
    end for
    for each $z \in De(y)$ do
        if $Edges(\delta_{yz}) \subseteq \mathcal{D}$ then
            update $\sigma'_{xz} = \sigma_{xz} - \varepsilon_{xy}\delta_{yz}$
        else
            remove $z$ from $G'$
        end if
    end for
    return $\langle G', \Sigma' \rangle$
end function
```

$$\sigma_{vw} = \sum_{\pi \in \mathcal{T}_{vw}} \pi(\Lambda, \mathcal{E})$$

At its core, identifiability of an edge λ in linear Gaussian SCM can be reduced to the problem of finding whether there exists a unique solution for λ in terms of covariances in the system of equations defined by the rules of path analysis [57], and knowledge of existing directed and bidirected effects.

With this in mind, we can prove PushForward.

Proof. Specified in the theorem is a covariance matrix Σ, a graph of the structural equations G, a set of known directed edges $\mathcal{D}$, and known bidirected edge ε_{xy}. The system of equations constraining values of structural parameters is

$$\sigma_{vw} = \sum_{\pi \in \mathcal{T}_{vw}} \pi(\Lambda, \mathcal{E}) \qquad \forall v, w \in G$$

We first look at σ_{xy}, and define a new known quantity σ'_{xy}:

$$\sigma_{xy} = \varepsilon_{xy} + \sum_{\pi \in \mathcal{T}_{xy} \setminus \{\varepsilon_{xy}\}} \pi(\Lambda, \mathcal{E})$$
$$\sigma'_{xy} = \sigma_{xy} - \varepsilon_{xy} = \sum_{\pi \in \mathcal{T}_{xy} \setminus \{\varepsilon_{xy}\}} \pi(\Lambda, \mathcal{E})$$

We also look at all descendants of y, $(z \in Z)$ where the directed paths from y to z (δ_{yz}) are made entirely of known edges $(Edges(\delta_{yz}) \subseteq \mathcal{D})$. We define

$$\delta_{ab} = \frac{1}{\varepsilon_a^2} \sum_{\pi \in \mathcal{T}_{xy}^{\rightarrow}} \pi(\Lambda, \mathcal{E})$$

where $\mathcal{T}_{xy}^{\rightarrow}$ represents the set of treks taking only directed edges from a to b: $a \rightarrow ... \rightarrow b$. For each such descendant of y, z, we define the quantity σ'_{xz}

$$\sigma_{xz} = \delta_{yz}\varepsilon_{xy} + \sum_{\pi \in \mathcal{T}_{xy} \setminus \mathcal{T}_{\varepsilon_{xy} yz}^{\rightarrow}} \pi(\Lambda, \mathcal{E})$$
$$\sigma'_{xz} = \sigma_{xz} - \delta_{yz}\varepsilon_{xy} = \sum_{\pi \in \mathcal{T}_{xy} \setminus \mathcal{T}_{\varepsilon_{xy} yz}^{\rightarrow}} \pi(\Lambda, \mathcal{E})$$

Here, we used $\mathcal{T}_{\varepsilon_{xy} yz}^{\rightarrow}$ to represent the treks starting from ε_{xy}, and continuing from y to x (half-treks from x to z using ε_{xy}). Finally, we define $\sigma'_{vw} = \sigma_{vw}$ for all other covariances between nodes a and b where both a and b are either non-descendants of y, or have their paths to y known.

This gives us a new system of equations in the original variables. All that remains to be shown is that an identified quantity λ'_{ab} in G' which contains a “pushed-forward” bidirected edge guarantees that the above-generated system of equations can be solved for the corresponding variable λ_{ab}.

As per the definition of G', it is identical to G, except:

1. the bidirected edge $x \leftrightarrow y$ is removed

2. $\forall c \in Ch(x)$, the edges $c \leftrightarrow y$ are added.

3. Descendants of y, z, where all edges of δ_{yz} are not known are removed

This new model G', with parameters Λ' and $\mathcal{E}'$ has system of equations:

$$\sigma''_{vw} = \sum_{\pi \in \mathcal{T}_{vw}} \pi(\Lambda', \mathcal{E}') \qquad \forall v, w \in G'$$

We compare this new system of equations to the modified equations of G.

- For all non-descendants of x or y, all covariance equations are identical (Both graphs have the same treks from non-descendants of x and y to all other nodes, and these covariances were not modified in the augmented equations).

- For all descendants of x, the modified equations for G have $\varepsilon_{xy}\lambda_{xc}$ wherever G' has ε'_{cy} when G does not have bidirected edge $c \leftrightarrow y$. If G already includes an ε_{cy}, then it has $(\varepsilon_{xy}\lambda_{xc} + \varepsilon_{cy})$ for each ε'_{cy}. This can be seen by comparing the treks available in the two models. We can create a map of treks in G to treks in G'. Treks not crossing the added/removed bidirected edges are identical. All that remains are treks crossing ε_{xy} in G, and ε'_{cy} in G'. Suppose we have a trek from a to b in G' $a \leftarrow ... \leftarrow c \leftrightarrow y \rightarrow ... \rightarrow b$, crossing the bidirected edge ε'_{cy}. The corresponding trek in G across ε_{cy}, if it exists, and the trek $a \leftarrow ... \leftarrow c \rightarrow x \leftrightarrow y \rightarrow ... \rightarrow b$ both map to it. Since we have a map from treks in G to all treks in G', which differs only in the specified spot, the equations are likewise identical save for the mapping difference.

- The covariances between x and the descendants of y and y have likewise identical equations. This is because the removed treks in the modified equations are the only possibilities including ε_{xy}, so all variables behave as if the bidirected edge did not exist. This can also be seen by recognizing that setting $\varepsilon_{xy} = 0$ would result in the same equation as removing all instances of the variable. Since the only treks from x which include ε_{xy} start by crossing $x \leftrightarrow y$, and continue on a directed path, removing all directed paths from y multiplied by ε_{xy} achieves the desired effect.

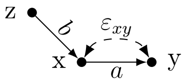
Finally, we notice that any algorithm for identifiability in this new model G' certifies that the system of equations can be uniquely solved for a given parameter, and the answer can be written in terms of Σ', which is computable given Σ and ε_{xy}. □

7.3.3 Identification, sensitivity analysis and Gröbner bases

Gröbner bases are a symbolic method of computer algebra used to solve systems of polynomial equations. [66] have shown that the identification (ID) problem in linear SCMs can be reduced to solving a system of polynomial equations and how Gröbner bases provide a complete solution.

In this section we will take a practical approach of showing how to set up the ID problem so it can be solved with Gröbner bases. We also show how to extend this to include sensitivity parameters, solving the problem of θ-identification. Our approach is based on [66]. For a basic understanding of Gröbner bases, please refer to [43].

Gröbner bases can be seen as an algorithm to do variable elimination in complex polynomial equations. Let us illustrate the variable elimination approach in the simple instrumental variable graph:



We can write the (normalized) covariance equations induced by the graph as follows:

$$\sigma_{xy} = a + \varepsilon_{xy}$$

$$\sigma_{zy} = b \times a$$

$$\sigma_{zx} = b$$

Given these equations, the goal is to solve for a in terms of the covariances of Σ only. Normally, one would approach this directly, by simply eliminating one variable at a time. For example, after eliminating b, we get:

$$\sigma_{xy} = a + \varepsilon_{xy}$$

$$\sigma_{zy} = \sigma_{zx} \times a$$

Next, we would eliminate ε_{xy}, by putting it in terms of a:

$$\varepsilon_{xy} = \sigma_{xy} - a$$

Then, we have a final equation just in terms a and the Σ. This equation can be solved for a, and depending on how many values of a satisfy the constraint, it give us our identification result (here, only $a = \frac{\sigma_{zy}}{\sigma_{zx}}$ is valid).

Gröbner bases perform an equivalent operation—they successively eliminate variables from the system of equations. In this situation, we want to eliminate ε_{xy} and b, leaving only a and the covariances. In SAGE [136], this reduces to the following code:

```
R.<a,b,epsilon_xy,sigma_zx,sigma_zy,sigma_xy>
                                = PolynomialRing(QQ)
Ideal(
    sigma_xy - (a+epsilon_xy),
    sigma_zy - (b*a),
    sigma_zx - (b)
).elimination_ideal([epsilon_xy,b]).groebner_basis()
```

If the result is a first degree polynomial in a, there is a single solution.

The extension of this method to the θ-identification problem entailed by sensitivity analysis is straightforward. As sensitivity parameters are treated like known variables, we simply do not eliminate them. In the above example, if we were to treat ε_{xy} as a sensitivity parameter, our code would be:

```
R.<a,b,epsilon_xy,sigma_zx,sigma_zy,sigma_xy>
                                = PolynomialRing(QQ)
Ideal(
    sigma_xy - (a+epsilon_xy),
    sigma_zy - (b*a),
    sigma_zx - (b)
).elimination_ideal([b]).groebner_basis()
```

with an identical interpretation: if the resulting polynomials in a, Σ and ε_{xy} are linear in a, we conclude that knowing the givens is sufficient to identify a.

Unfortunately, despite the completeness of this approach, Gröbner bases are doubly-exponential in the number of variables, and in this case *each edge* corresponds to a variable [14]. This limits the practical solvable graph size to 4 or 5 nodes [57, 66]. Our own experiments hit upon the same limitation, with attempted computations on 5-node graphs sometimes taking several days for identifying single edges, despite using an optimized representation of the equations [57].

7.3.4 Detailed description of computational experiments

In this section, we provide a detailed description of our computational experiments, including pseudocode and additional tests. Our computational experiments have two main goals.

First, they aim to empirically verify the generality of our constrained identification algorithm cID, by comparing our results to the ground truth obtained via computer algebra.

Second, note that cID has three separate components:

1. The qID algorithm [28], which we use both for the identification of directed edges, and for incorporating constraints on directed edges that can be used as sensitivity parameters;

2. The graphical manipulations performed by PushForward, which we use to incorporate constraints on bidirected edges; and,

3. The order in which to perform the graphical manipulation of PushForward. In cID we chose to perform a topological ordering as described in Algorithm 1.

Thus, our computational experiments also aim to disentangle the contributions of each of those components to our results.

Solving all 3 and 4-Node sensitivity queries

Our computational experiments rely on the ability to find ground-truth answers to the question of whether a target coefficient λ_{ab} is θ-identifiable in a given graph G (this is defined to be a *query*). As explained in Section 7.3.3, these ground truth answers can be obtained with algebraic methods, more precisely using Gröbner bases [66].

For 3-node models we have 50 connected graphs with 720 possible queries; for 4-node models, we have 3,745 connected graphs and 1,059,156 possible queries. Note that, for 5-node models, we have 1,016,317 connected graphs and 11,615,669,904 possible queries. As mentioned in Section 7.3.3, ground-truth computations using computer algebra can take hours (or sometimes days) for a *single* 5-node graph, rendering an exhaustive study of sensitivity queries in 5-node models impractical.

We have thus performed an exhaustive computation of the ground truth answer of all possible queries in 3 and 4 node models via computer algebra using SAGE [136]. These results give us a list stating for every graph G, every edge λ_{ab}, and *all possible subsets* of directed and bidirected edges used as sensitivity parameters θ, whether λ_{ab} can be uniquely computed from Σ and θ.

Our main interest lies on those queries that can be identified only when $\theta \neq \emptyset$ (we call this a sensitivity query)—in other words, we do not consider those edges that can be identified from Σ alone, since in these cases the parameter is identifiable and a sensitivity analysis would not be needed. The ground truth numbers of all θ-identifiable queries only when $\theta \neq \emptyset$ are 320 for 3-node models and 578,858 for 4-node models.

Our exhaustive computations also allow us to see how many sensitivity queries can be solved using *only subsets of directed edges* or *only subsets of bidirected edges* as sensitivity parameters. The decomposition then becomes the following. For 3-node models, there are 19 sensitivity queries that can be solved using only subsets of directed edges as sensitivity parameters, 109 using only subsets of bidirected edges, and, as before, 320 total queries which are solvable using an arbitrary combination of both. For 4-node models, these numbers increase to 15,740, 52,016 and 578,858 respectively. These numbers reveal that incorporating constraints on bidirected edges is an essential step for deriving sensitivity curves.

Comparing qID and cID to ground-truth answers

Once we have obtained ground-truth answers to all queries in 3 and 4-node models, we run both the qID as well as the cID algorithm for each of those queries and check whether they can correctly decide whether θ is an admissible set of sensitivity parameters for λ_{ab} in G (and thus able to provide a sensitivity curve). This comparison gives us the numbers we have presented in the main text in Table 4.1.

		3 Nodes			4 Nodes		
PF order	ID Alg. directed edges	*Directed*	*Bidirected*	*Both*	*Directed*	*Bidirected*	*Both*
none	qID	19	-	68	14,952	-	170,304
PFo	qID	19	101	304	14,952	43,526	505,076
PFs	qID	19	105	308	14,952	46,630	517,036
PFr	qID	19	109	320	14,952	50,708	555,758
PFt	**qID**	19	109	320	14,952	50,708	555,758
none	Complete	19	-	68	15,740	-	177,216
PFo	Complete	19	101	304	15,740	44,680	524,846
PFs	Complete	19	105	308	15,740	47,962	538,332
PFr	Complete	19	109	320	15,740	51,992	578,758
PFt	Complete	19	109	320	15,740	51,992	578,758
Ground Truth		19	109	320	15,740	52,016	578,858

Table 7.1: Number of θ-identifiable queries (only when $\theta \neq \emptyset$) per type of sensitivity parameters θ, using different ordering methods for PushForward and different ID algorithm for the directed edges.

Note: Ground Truth is computed using Gröbner bases. The first column defines the ordering method of PushForward used for incorporating constraints on bidirected edges—this is passed as the argument PFOrder in the general function cID*. The second column refers to the identification algorithm used for directed edges—this is passed as the argument IDmethod in the general function cID*. "Complete" means we used Gröbner bases to simulate a complete ID algorithm for *directed edges* running inside cID*. Note the first row corresponds to qID and the boldfaced row corresponds to cID as presented in the main text applying PushForward in topological ordering. These two rows are the ones presented in Table 4.1 of the main text. Pseudocode for computing these numbers is given in Algorithm 4.

Alternative ordering methods for PushForward

In the main text, the cID algorithm applies PushForward in a topological ordering for processing multiple bidirected edges. The method does not perform all possible graphical manipulations, and as such, a valid concern is that it might be less capable than a more general search. Another interesting question is to check whether simpler methods would perform as well as the current cID implementation. To tackle these questions, we tested additional ordering methods for handling multiple bidirected edges.

For simplicity of exposition, the cID algorithm in the main text has the ordering method embedded in the pseudocode itself. For the purposes of this section, however, it is conceptually easier to create a meta algorithm that repeats the following process: (i) first it creates a collection of valid modified graphs $\mathcal{G}$ applying PushForward according to some ordering method; then, (ii) it applies an identification algorithm to each of those modified graphs. This is given in Algorithm 3, which we call cID*.

Algorithm 3 Meta constrained ID algorithm.

```
function cID*($G, \Sigma, \mathcal{B}, \mathcal{D}$, PForder, IDmethod)
    repeat
        $\mathcal{G} \leftarrow$ PForder($G, \Sigma, \mathcal{B}, \mathcal{D}$)
        for $\langle G', \Sigma' \rangle \in \mathcal{G}$ do
            $\mathcal{D} \leftarrow \mathcal{D} \cup$ IDmethod($G', \Sigma', \mathcal{D}$)
        end for
    until all directed edges have been identified or no edge has been identified in the last iteration
    return $\mathcal{D}$
end function
```

In Algorithm 3, the argument PForder represents a function that takes as inputs a graph G, a covariance matrix Σ, a set of known bidirected edges $\mathcal{B}$ and a set of known directed edges $\mathcal{D}$. It then returns a *collection* $\mathcal{G}$ of valid modified models $\langle G', \Sigma' \rangle$ by iteratively applying PushForward following a particular ordering method (for example, topological ordering). The argument IDmethod refers to an identification method for directed edges (for instance, qID). It is a function that takes as inputs a graph G, a covariance matrix Σ and a set of known directed edges $\mathcal{D}$ and it returns the new set of known directed edges.

We can now create different functions for different ordering methods. For instance, the function PFt described in Algorithm 7 applies PushForward in topological ordering (as embedded in Algorithm 1 of the main text) and returns all valid modified graphs. We now define three additional ordering methods.

- PFo described in Algorithm 5. This function pushes forward each bidirected edge only once, considering the original graph. This method is the simplest application of PushForward, and serves as a base of comparison to assess the gains of more elaborate methods.

- PFs described in Algorithm 6. This function tries to apply PushForward once to all subsets of bidirected edges connected to each end node. This procedure has exponential computational complexity.

- PFr described in Algorithm 8. This function recursively tries every possible combination of applying PushForward for each bidirected edge connected to the same end node (it tries each subset once, and of those that can be pushed forward again, tries each subset, and so on). This procedure has doubly exponential computational complexity.

All these function return a collection $\mathcal{G}$ of valid modified graphs, and can be used as the PForder argument in the cID* function. Of these methdos, PFr is arguably the most important for comparison with our current implementation of topological ordering. The results are shown in the first half of Table 7.1, which compares cID* using the same ID method for directed edges (qID) but different ordering methods for applying PushForward. Our preferred version, which was presented in the main text as cID, corresponds to the boldfaced row with ordering method PFt and ID method qID. As we can see, topological ordering performs as well as the brute-force recursive search of all subsets performed by PFr, which has doubly exponential computational complexity.

Disentangling PF and qID

Finally, the incompleteness of cID can stem from two sources: limitations of the graphical manipulations performed by PushForward or the incompleteness of the identification algorithm for directed edges, qID. Separating the two can help guide efforts for future research. To achieve that, we used algebraic methods to simulate how cID would have performed if it had access to a complete identification algorithm for directed edges instead of qID.

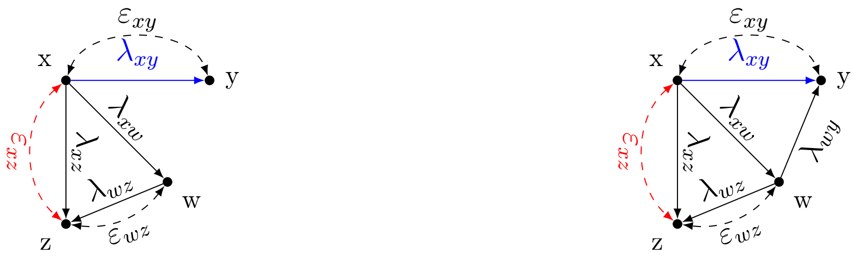
More precisely, we use Gröbner bases as our ID algorithm for directed edges (IDmethod) in cID*, where, just like qID, Gröbner bases only have access to constraints on bidirected edges via the graphical manipulation performed by PushForward. That is, Gröbner bases is dealing with the problem as if it were a "vanilla" identification problem, not explicitly knowing that the bidirected edge is fixed. The results can be seen in the second half of Table 7.1. The last row indicates, for instance, that incorporating constraints on bidirected edges using PushForward in topological order, in combination with a complete identification algorithm for directed edges, would have identified over 99.99% of 4-node sensitivity queries.

This suggests that: (i) the main bottleneck of the current implementation of cID is qID itself; (ii) PushForward with topological ordering is an efficient procedure for dealing with bidirected edges that can reap the benefits of improved identification algorithms.

7.3.5 The missed cases

As discussed in the previous section, PushForward in topological order, in combination with a complete identification algorithm for *directed edges*, would have identified over 99.99% of all 4-node sensitivity queries. In this section we briefly discuss some of the missed cases, which may provide guidance for further improvements of the cID algorithm. We also provide all the missed cases for those interested in exploring them further (Tables 7.2 and 7.3).

When iterating over modified graphs, the cID algorithm feeds its next iteration *only* identification results for *direct effects* (single coefficients), not of path specific effects or total



(a)

(b)

Figure 7.2: Examples of missed cases using PushForward with a complete identification algorithm of *directed edges*. In both examples, λ_{xy} is ε_{zx}-identifiable, but the algorithm fails due to lack of exploitation of identified total effects. In example 7.2a, it turns out a simple marginalization of w suffices for the ε_{zx}-identification of λ_{xy} using the current implementation of the cID algorithm. However, marginalization alone is often not enough, as shown in example 7.2b.

effects (sums of products of coefficients), which may nevertheless be identified. Figure 7.2 shows two simple examples that illustrates how not exploiting the knowledge of known total effects can result in a failure of identification.

Let us start with Figure 7.2a. In this example, our task is to find a sensitivity curve for λ_{xy} in terms of ε_{zx}. First note that z is not a valid instrument for λ_{xy} since it is a descendant of x. However, pushing forward ε_{zx} allows us to identify the *total effect* of x on z. This, in turn, permits the creation of the auxiliary variable $z^* = z - (\lambda_{xz} + \lambda_{xw}\lambda_{wz})x$ which is now a valid instrument for λ_{xy}. In the example of Figure 7.2a, it turns out a simpler solution would also suffice—marginalizing w. Note the marginalized DAG results in a simple three node model which can be solved by the current implementation of cID. Nevertheless, marginalization by itself may not always be sufficient, as a simple variation of this very example shows (Figure 7.2b).

In sum, θ-identification in these cases require systematically exploiting known total effects (for instance, creating AVs subtracting out total effects) or known path-specific effects, a task which still does not have a satisfactory solution in the literature. A final interesting (and challenging) example in which cID failed to find the sensitivity curve is shown in Figure 7.3.

Algorithm 4 Pseudocode for checking the performance of cID* with different PushForward orders and different ID algorithm for directed edges. In the code, IdentifyDirectedEdges and IsIdentified are computed using computer algebra (Gröbner bases), and give the ground-truth values.

```
initialize Total $\leftarrow 0$
initialize PFtotal $\leftarrow 0$
$\mathcal{S} \leftarrow$ set of all possible connected DAGs, with all combinations of directed and bidirected edges.
for each graph $\langle G, \Sigma \rangle \in \mathcal{S}$ do
    IDedges $\leftarrow$ IdentifyDirectedEdges($\mathcal{G}$)
    for all $(x \to y) \in \mathcal{G}$ where $(x \to y) \notin$ IDedges do
        $\mathcal{SPS} \leftarrow$ All subsets of directed and bidirected edges of $G$ which do not contain $(x \to y)$
        for each set $\langle \mathcal{D}, \mathcal{B} \rangle \in \mathcal{SPS}$ do
            if IsIdentified($\mathcal{G}, x \to y, \mathcal{D}, \mathcal{B}$) then
                Total $\leftarrow$ Total + 1
                if $(x \to y) \in$ cID*($G, \Sigma, \mathcal{D}, \mathcal{B}$, PForder, IDmethod) then
                    PFTotal $\leftarrow$ PFTotal + 1
                end if
            end if
        end for
    end for
end for
return $\frac{PFtotal}{Total}$
```

Algorithm 5 Push forward each bidirected edge once.

```
function PFo($G, \Sigma, \mathcal{B}, \mathcal{D}$)
    let $\mathcal{B}_y$ represent subset of $\mathcal{B}$ where all edges have $y$ as end point ($B_y = \{(x \leftrightarrow y) \in \mathcal{B}, \forall x\}$)
    $\mathcal{G} \leftarrow \{(G, \Sigma)\}$
    for each node $y \in G$ do
        for bidirected edge $\varepsilon_{xy} \in \mathcal{B}_y$ do
            if $x \notin$ De($y$) or $\delta_{yx} \in \mathcal{D}$ then
                add PF($G, \Sigma, \mathcal{D}, \varepsilon_{xy}, x$) to $\mathcal{G}$
            end if
        end for
    end for
    return $\mathcal{G}$
end function
```

Algorithm 6 Push forward all subsets once.

```
function PFs($G, \Sigma, \mathcal{B}, \mathcal{D}$)
    let $\mathcal{B}_y$ represent subset of $\mathcal{B}$ where all edges have $y$ as end point ($B_y = \{(x \leftrightarrow y) \in \mathcal{B}, \forall x\}$)
    $\mathcal{G} \leftarrow \{(G, \Sigma)\}$
    for each node $y$ do
        for each $B'_y \subseteq B_y$ do
            $\langle G', \Sigma' \rangle \leftarrow \langle G, \Sigma \rangle$
            for each $\varepsilon_{xy} \in B'_y$ do
                if $x \notin$ De($y$) or $\delta_{yx} \in \mathcal{D}$ then
                    $\langle G', \Sigma' \rangle \leftarrow$PF($G', \Sigma', \mathcal{D}, \varepsilon_{xy}, x$)
                    for all $z \in$ Ch($x$) do
                        if $\lambda_{xz} \notin \mathcal{D}$ then
                            remove $\varepsilon_{zy}$ from $\mathcal{B}'_y$ if it was not yet processed.
                        end if
                    end for
                end if
            end for
            add $\langle G', \Sigma' \rangle$ to $\mathcal{G}$
        end for
    end for
    return $\mathcal{G}$
end function
```

Algorithm 7 Push forward in topological order.

```
function PFT($G, \Sigma, \mathcal{B}, \mathcal{D}$)
  let $\mathcal{B}_y$ represent subset of $\mathcal{B}$ where all edges have $y$ as end point ($B_y = \{(x \leftrightarrow y) \in \mathcal{B}, \forall x\}$)
  $\mathcal{G} \leftarrow \{(G, \Sigma)\}$
  for each node $y$ do
    $\langle G', \Sigma' \rangle \leftarrow \langle G, \Sigma \rangle$
    for each $\varepsilon_{xy} \in B_y$ in topological order on $x$ do
      if $x \notin \textsc{De}(y)$ or $\delta_{yx} \in \mathcal{D}$ then
        $\langle G', \Sigma' \rangle \leftarrow$PF$(G', \Sigma', \mathcal{D}, \varepsilon_{xy}, x)$
        add $\langle G', \Sigma' \rangle$ to $\mathcal{G}$
        for all $z \in \textsc{Ch}(x)$ do
          if  $\lambda_{xz} \notin \mathcal{D}$ then
            remove $\varepsilon_{zy}$ from $\mathcal{B}_y$ if it was not yet processed.
          else
            add $\varepsilon_{zy}$ to $\mathcal{B}_y$
          end if
        end for
      end if
    end for
  end for
  return $\mathcal{G} \cup$ PFO$(G, \Sigma, \mathcal{B}, \mathcal{D})$
end function
```

Algorithm 8 Push forward all subsets recursively.

```
function $\textsc{PFr}(G, \Sigma, \mathcal{B}, \mathcal{D})$
  let $\mathcal{B}_y$ represent subset of $\mathcal{B}$ where all edges have $y$ as end point ($B_y = \{(x \leftrightarrow y) \in \mathcal{B}, \forall x\}$)
  initialize $\mathcal{G} \leftarrow \{(G, \Sigma, \emptyset)\}$
  for each node $y$ do
    $PushSets \leftarrow \{\langle G, \Sigma, B'_y \rangle$ for all $B'_y \subseteq B_y\}$
    while $PushSets$ not empty do
      pop $\langle G', \Sigma', B'_y \rangle$ from $PushSets$
      $PushAgain \leftarrow \{\}$
      for each $\varepsilon_{xy} \in B'_y$ do
        if $x \notin \textsc{De}(y)$ or $\delta_{yx} \in \mathcal{D}$ then
          $\langle G', \Sigma' \rangle \leftarrow \mathrm{PF}(G', \Sigma', \mathcal{D}, \varepsilon_{xy}, x)$
          for all $z \in \textsc{Ch}(x)$ do
            if  $\lambda_{xz} \notin \mathcal{D}$ then
              remove $\varepsilon_{zy}$ from $\mathcal{B}'_y$ if it was not yet processed.
            else
              add $\varepsilon_{zy}$ to $PushAgain$
            end if
          end for
        end if
      end for
      add $\langle G', \Sigma' \rangle$ to $\mathcal{G}$
      for all$B''_y \subseteq PushAgain$ do
        add $\langle G', \Sigma', B''_y \rangle$ to $PushSets$
      end for
    end while
  end for
  return $\mathcal{G}$
end function
```

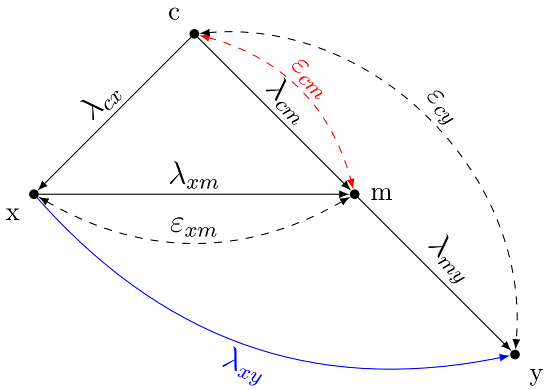


Figure 7.3: An interesting missed case example. Here λ_{xy} is ε_{cm}-identifiable. All examples can be found in Tables 7.2 and 7.3.

	Graph	Target Quantity	Sensitivity Parameters
1	1→2 1→3 1→4 2→4 1↔3 1↔4 2↔4	1→3	1↔4
2	1→2 1→3 1→4 2→4 1↔3 1↔4 2↔4	1→3	1↔4 1→2
3	1→2 1→3 1→4 2→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4
4	1→2 1→3 1→4 2→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 1→2
5	1→2 1→3 1→4 2→4 3→4 1↔3 1↔4 2↔4	1→3	1↔4 3→4
6	1→2 1→3 1→4 2→4 3→4 1↔3 1↔4 2↔4	1→3	1↔4 1→2 3→4
7	1→2 1→3 1→4 2→4 3→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 3→4
8	1→2 1→3 1→4 2→4 3→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 1→2 3→4
9	1→2 1→3 2→3 1→4 2→4 1↔3 1↔4 2↔4	1→3	1↔4
10	1→2 1→3 2→3 1→4 2→4 1↔3 1↔4 2↔4	1→3	1↔4 2→3
11	1→2 1→3 2→3 1→4 2→4 1↔3 1↔4 2↔4	1→3	1↔4 1→2
12	1→2 1→3 2→3 1→4 2→4 1↔3 1↔4 2↔4	1→3	1↔4 1→2 2→3
13	1→2 1→3 2→3 1→4 2→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4
14	1→2 1→3 2→3 1→4 2→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 2→3
15	1→2 1→3 2→3 1→4 2→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 1→2
16	1→2 1→3 2→3 1→4 2→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 1→2 2→3
17	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 1↔4 2↔4	1→3	1↔4 3→4
18	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 1↔4 2↔4	1→3	1↔4 3→4 1→2
19	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 1↔4 2↔4	1→3	1↔4 2→3 3→4
20	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 1↔4 2↔4	1→3	1↔4 2→3 3→4 1→2
21	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 3→4
22	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 3→4 1→2
23	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 2→3 3→4
24	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 1↔4 2↔4 3↔4	1→3	1↔4 3↔4 2→3 3→4 1→2
25	1→2 1→3 2→3 2→4 3→4 1↔3 2↔3 1↔4	2→4	1↔3
26	1→2 1→3 2→3 2→4 3→4 1↔3 2↔3 1↔4	2→4	1↔3 1→2
27	1→2 1→3 2→3 2→4 3→4 1↔3 2↔3 1↔4	3→4	1↔3
28	1→2 1→3 2→3 2→4 3→4 1↔3 2↔3 1↔4	3→4	1↔3 1→2
29	1→2 1→3 2→3 2→4 3→4 1↔3 2↔3 1↔4 3↔4	2→4	1↔3 3↔4
30	1→2 1→3 2→3 2→4 3→4 1↔3 2↔3 1↔4 3↔4	2→4	1↔3 3↔4 1→2
31	1→2 1→3 2→3 2→4 3→4 1↔3 2↔3 1↔4 3↔4	3→4	1↔3 3↔4
32	1→2 1→3 2→3 2→4 3→4 1↔3 2↔3 1↔4 3↔4	3→4	1↔3 3↔4 1→2
33	1→2 1→3 2→3 1→4 1↔3 2↔3 1↔4	1→4	1↔3
34	1→2 1→3 2→3 1→4 1↔3 2↔3 1↔4	1→4	1↔3 1→2
35	1→2 1→3 2→3 1→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4
36	1→2 1→3 2→3 1→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 1→2
37	1→2 1→3 2→3 1→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3
38	1→2 1→3 2→3 1→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 3→4
39	1→2 1→3 2→3 1→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 1→2
40	1→2 1→3 2→3 1→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 1→2 3→4
41	1→2 1→3 2→3 1→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4
42	1→2 1→3 2→3 1→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 3→4
43	1→2 1→3 2→3 1→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 1→2
44	1→2 1→3 2→3 1→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 1→2 3→4
45	1→2 1→3 2→3 1→4 2→4 1↔3 2↔3 1↔4	1→4	1↔3
46	1→2 1→3 2→3 1→4 2→4 1↔3 2↔3 1↔4	1→4	1↔3 2→4
47	1→2 1→3 2→3 1→4 2→4 1↔3 2↔3 1↔4	1→4	1↔3 1→2
48	1→2 1→3 2→3 1→4 2→4 1↔3 2↔3 1↔4	1→4	1↔3 1→2 2→4
49	1→2 1→3 2→3 1→4 2→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4
50	1→2 1→3 2→3 1→4 2→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 2→4

Table 7.2: Missed sensitivity queries of PushForward in topological order, in combination with a complete identification algorithm for *directed edges*. Part 1.

	Graph	Target Quantity	Sensitivity Parameters
51	1→2 1→3 2→3 1→4 2→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 1→2
52	1→2 1→3 2→3 1→4 2→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 1→2 2→4
53	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 3→4
54	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 3→4 1→2
55	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 2→4
56	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 1→2 2→4
57	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 3→4 2→4
58	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 3→4 1→2 2→4
59	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 1↔4
60	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	1→4	1↔3 1↔4 1→2
61	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	2→4	1↔3 1→4
62	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	2→4	1↔3 1→2 1→4
63	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	2→4	1↔3 1↔4
64	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	2→4	1↔3 1↔4 1→2
65	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	3→4	1↔3 1→4
66	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	3→4	1↔3 1→2 1→4
67	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	3→4	1↔3 1↔4
68	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4	3→4	1↔3 1↔4 1→2
69	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 3→4
70	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 3→4 1→2
71	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 2→4
72	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 1→2 2→4
73	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 3→4 2→4
74	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 3↔4 3→4 1→2 2→4
75	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 1↔4 3↔4
76	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	1→4	1↔3 1↔4 3↔4 1→2
77	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	2→4	1↔3 3↔4 1→4
78	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	2→4	1↔3 3↔4 1→2 1→4
79	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	2→4	1↔3 1↔4 3↔4
80	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	2→4	1↔3 1↔4 3↔4 1→2
81	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	3→4	1↔3 3↔4 1→4
82	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	3→4	1↔3 3↔4 1→2 1→4
83	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	3→4	1↔3 1↔4 3↔4
84	1→2 1→3 2→3 1→4 2→4 3→4 1↔3 2↔3 1↔4 3↔4	3→4	1↔3 1↔4 3↔4 1→2
85	1→2 2→3 2→4 3→4 1↔2 1↔4 3↔4	1→2	1↔4
86	1→2 2→3 2→4 3→4 1↔2 1↔4 3↔4	1→2	1↔4 2→3
87	1→2 2→3 2→4 3→4 1↔2 1↔4 2↔4 3↔4	1→2	1↔4 2↔4
88	1→2 2→3 2→4 3→4 1↔2 1↔4 2↔4 3↔4	1→2	1↔4 2↔4 2→3
89	1→2 2→3 1→4 2→4 3→4 1↔2 1↔4 3↔4	1→2	1↔4 1→4
90	1→2 2→3 1→4 2→4 3→4 1↔2 1↔4 3↔4	1→2	1↔4 2→3 1→4
91	1→2 2→3 1→4 2→4 3→4 1↔2 1↔4 2↔4 3↔4	1→2	1↔4 2↔4 1→4
92	1→2 2→3 1→4 2→4 3→4 1↔2 1↔4 2↔4 3↔4	1→2	1↔4 2↔4 2→3 1→4
93	1→2 1→3 1→4 3→4 1↔2 1↔4 3↔4	1→2	1↔4
94	1→2 1→3 1→4 3→4 1↔2 1↔4 3↔4	1→2	1↔4 1→3
95	1→2 1→3 1→4 3→4 1↔2 1↔4 2↔4 3↔4	1→2	1↔4 2↔4
96	1→2 1→3 1→4 3→4 1↔2 1↔4 2↔4 3↔4	1→2	1↔4 2↔4 1→3
97	1→2 1→3 1→4 2→4 3→4 1↔2 1↔4 3↔4	1→2	1↔4 2→4
98	1→2 1→3 1→4 2→4 3→4 1↔2 1↔4 3↔4	1→2	1↔4 1→3 2→4
99	1→2 1→3 1→4 2→4 3→4 1↔2 1↔4 2↔4 3↔4	1→2	1↔4 2↔4 2→4
100	1→2 1→3 1→4 2→4 3→4 1↔2 1↔4 2↔4 3↔4	1→2	1↔4 2↔4 1→3 2→4

Table 7.3: Missed sensitivity queries of PushForward in topological order, in combination with a complete identification algorithm for *directed edges*. Part 2.

7.4 Appendix for Chapter 5

7.4.1 An example with continuous treatment

Here we provide a simple example in which, although the treatment variable is continuous, the relevant dependencies among potential outcomes are still amenable to graphical representation. Suppose we have the same selection diagram as in Figure 5.2b, but now let X, B, and H all be continuous variables. Next, consider the following functional specification for the structural equation of Y,

$$Y = I(H > 0) \vee I(X \times B > 0) \tag{7.130}$$

Where $I(\cdot)$ denotes the indicator function. Now note from Equation 7.130 we can derive the potential outcomes $Y_0 = I(H > 0)$ for $x = 0$, and, $Y_x = I(H > 0) \vee I(xB > 0) = Y_0 \vee I(xB > 0)$, for $x \neq 0$. We can thus draw the same modified selection diagram as in Figure 5.3, but now replacing Y_1 with Y_x, leading to the conclusion that $Y_x \perp\!\!\!\perp S \mid Y_0$, for all $x \neq 0$.

7.4.2 Proofs

7.4.2.1 Bounds with a single source population

Here we show how to obtain the bounds of Theorem 2. To simplify notation, let $P_{ij} := P(Y_i = j)$, $P^*_{ij} := P^*(Y_i = j)$, $\text{PS}_{10} := P^*(Y_1 = 0|Y_0 = 1) = P(Y_1 = 0|Y_0 = 1)$ and $\text{PS}_{01} = P^*(Y_1 = 1|Y_0 = 0) = P(Y_1 = 1|Y_0 = 0)$. The target function to be optimized is P^*_{11}, which can be written as,

$$P^*_{11} = (1 - \text{PS}_{10})P^*_{01} + \text{PS}_{01}(1 - P^*_{01}) \tag{7.131}$$

Our goal is to pick PS_{10} and PS_{01} such that it maximizes (or minimizes) Equation 7.131 subject to the following constraints: (i) PS_{10} and PS_{01} need to be between zero and one (since PS_{10} and PS_{01} need to be valid probabilities); and, (ii) PS_{10} and PS_{01} must conform to the observed results of the trial in the source domain, that is, $P_{11} = (1-\mathrm{PS}_{10})P_{01} + \mathrm{PS}_{01}(1-P_{01})$. Thus, our optimization problem is,

$$\begin{aligned} \max_{\mathrm{PS}_{10},\mathrm{PS}_{01}} \quad & P^*_{11} = (1-\mathrm{PS}_{10})P^*_{01} + \mathrm{PS}_{01}(1-P^*_{01}) \\ \text{s.t.} \quad & P_{11} = (1-\mathrm{PS}_{10})P_{01} + \mathrm{PS}_{01}(1-P_{01}) \\ \text{and} \quad & 0 \le \mathrm{PS}_{10} \le 1,\ 0 \le \mathrm{PS}_{01} \le 1 \end{aligned}$$

To simplify the problem, we can use the equality constraint $P_{11} = (1-\mathrm{PS}_{10})P_{01} + \mathrm{PS}_{01}(1-P_{01})$ to eliminate one of the variables. For instance, writing PS_{10} in terms of PS_{01} gives us,

$$1 - \mathrm{PS}_{10} = \frac{P_{11} - \mathrm{PS}_{01}(1-P_{01})}{P_{01}} \tag{7.132}$$

Which results in a new target function,

$$\begin{aligned} P^*_{11} &= (1-\mathrm{PS}_{10})P^*_{01} + \mathrm{PS}_{01}(1-P^*_{01}) && (7.133) \\ &= \left(\frac{P_{11} - \mathrm{PS}_{01}(1-P_{01})}{P_{01}}\right) P^*_{01} + \mathrm{PS}_{01}(1-P^*_{01}) && (7.134) \\ &= \left(\frac{P_{11}}{P_{01}}\right) P^*_{01} + \left(\frac{P_{01}-P^*_{01}}{P_{01}}\right) \mathrm{PS}_{01} && (7.135) \\ &= RR \times P^*_{01} + \left(\frac{P_{01}-P^*_{01}}{P_{01}}\right) \mathrm{PS}_{01} && (7.136) \end{aligned}$$

Where $RR = \frac{P_{11}}{P_{01}}$ is the causal *risk-ratio* in the trial of the source domain Π. Since $0 \le (1-\mathrm{PS}_{10}) \le 1$, the substitution also results in additional constraints on PS_{01},

$$\frac{P_{11} - P_{01}}{1 - P_{01}} \le \mathrm{PS}_{01} \le \frac{P_{11}}{1-P_{01}} \tag{7.137}$$

Thus, define the lower and upper bounds on PS_{01} as

$$\mathrm{PS}^L_{01} = \max\left\{0, \frac{P_{11}-P_{01}}{1-P_{01}}\right\}, \qquad \mathrm{PS}^U_{01} = \min\left\{\frac{P_{11}}{1-P_{01}}, 1\right\}$$

Our new maximization problem can be written as,

$$\max_{\mathrm{PS}_{01}} \quad RR \times P^*_{01} + \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\mathrm{PS}_{01} \qquad \text{s.t.} \quad \mathrm{PS}^L_{01} \leq \mathrm{PS}_{01} \leq \mathrm{PS}^U_{01} \tag{7.138}$$

Since the target function is linear, the maximum occurs at the extreme points of PS_{01}. The same reasoning holds for the minimization problem. Thus, we have that,

$$P^{*L}_{11} \leq P^*_{11} \leq P^{*U}_{11}$$

Where,

$$P^{*L}_{11} = RR \times P^*_{01} + \min\left\{\left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\mathrm{PS}^L_{01},\ \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\mathrm{PS}^U_{01}\right\}$$

and

$$P^{*U}_{11} = RR \times P^*_{01} + \max\left\{\left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\mathrm{PS}^L_{01},\ \left(\frac{P_{01} - P^*_{01}}{P_{01}}\right)\mathrm{PS}^U_{01}\right\}$$

7.4.2.2 Informativeness of the bounds

We now derive the width of the bounds for P^*_{11} for the case when the bounds for PS_{01} do not reach 0 nor 1 (this will happen when both $P_{11} > P_{01}$ and $P_{11} < 1 - P_{01}$). Define the width W of the bounds as the difference between the upper and lower bound of P^*_{11}, that is,

$$W = P^{*U}_{11} - P^{*L}_{11}$$

Expanding the terms we obtain,

$$
\begin{aligned}
W &= P_{11}^{*U} - P_{11}^{*L} && (7.139)\\
&= \left|\left(\frac{P_{01} - P_{01}^{*}}{P_{01}}\right)\mathrm{PS}_{01}^{U} - \left(\frac{P_{01} - P_{01}^{*}}{P_{01}}\right)\mathrm{PS}_{01}^{L}\right| && (7.140)\\
&= \frac{|P_{01} - P_{01}^{*}|}{P_{01}} \times \left(PS_{01}^{U} - \mathrm{PS}_{01}^{L}\right) && (7.141)\\
&= \frac{|P_{01} - P_{01}^{*}|}{P_{01}} \times \frac{P_{01}}{1 - P_{01}} && (7.142)\\
&= \frac{|P_{01} - P_{01}^{*}|}{1 - P_{01}} && (7.143)
\end{aligned}
$$

Thus, when the bounds for PS_{01} are "interior," the informativeness of the bounds depends only on P_{01} and P_{01}^{*}. Moreover, even if the bounds for PS_{01} are "wide," the bounds for P_{11}^{*} may be "narrow," provided the baseline risks of the source and target population are close enough.

7.4.2.3 Identification with multiple source domains

We now show how to obtain the identification results of Theorem 3 and 4. Consider two source populations Π^{a} and Π^{b}. Again, to simplify notation, let $P_{ij}^{a} := P^{a}(Y_i = j)$, $P_{ij}^{b} := P^{a}(Y_i = j)$, $\mathrm{PS}_{10} := P^{a}(Y_1 = 0|Y_0 = 1) = P^{b}(Y_1 = 0|Y_0 = 1) = P^{*}(Y_1 = 0|Y_0 = 1)$ and $\mathrm{PS}_{01} := P^{a}(Y_1 = 1|Y_0 = 0) = P^{b}(Y_1 = 1|Y_0 = 0) = P^{*}(Y_1 = 1|Y_0 = 0)$.

First note that PS_{10} and PS_{01} are identified from the experimental data in Π^{a} and Π^{b}. Using the law of total probability for P_{11}^{a} and P_{11}^{b} write,

$$
\begin{aligned}
P_{11}^{a} &= (1 - \mathrm{PS}_{10}) \times P_{01}^{a} + \mathrm{PS}_{01} \times P_{00}^{a} && (7.144)\\
P_{11}^{b} &= (1 - \mathrm{PS}_{10}) \times P_{01}^{b} + \mathrm{PS}_{01} \times P_{00}^{b} && (7.145)
\end{aligned}
$$

We thus have a system of two equations and two unknowns,

$$\begin{bmatrix} P_{01}^a & P_{00}^a \\ P_{01}^b & P_{00}^b \end{bmatrix} \begin{bmatrix} (1 - \mathrm{PS}_{10}) \\ \mathrm{PS}_{01} \end{bmatrix} = \begin{bmatrix} P_{11}^a \\ P_{11}^b \end{bmatrix} \tag{7.146}$$

Yielding the solution,

$$\begin{bmatrix} (1 - \mathrm{PS}_{10}) \\ \mathrm{PS}_{01} \end{bmatrix} = \frac{1}{P_{01}^a P_{00}^b - P_{01}^b P_{00}^a} \times \begin{bmatrix} P_{00}^b & -P_{00}^a \\ -P_{01}^b & P_{01}^a \end{bmatrix} \begin{bmatrix} P_{11}^a \\ P_{11}^b \end{bmatrix} \tag{7.147}$$

Which amounts to:

$$\mathrm{PS}_{10} = 1 - \frac{P_{11}^a P_{00}^b - P_{11}^b P_{00}^a}{P_{01}^a P_{00}^b - P_{01}^b P_{00}^a} \tag{7.148}$$

$$\mathrm{PS}_{01} = \frac{P_{11}^b P_{01}^a - P_{11}^a P_{01}^b}{P_{01}^a P_{00}^b - P_{01}^b P_{00}^a} \tag{7.149}$$

All values of the RHS can be computed from the experimental data of Π^a and Π^b. Note that, since PS_{10} and PS_{01} must be between 0 and 1, not all solutions are valid. Therefore, two domains already entail some testable implications—if either PS_{10} and PS_{01} are not valid probabilities, this means that the assumption that the probabilities of sufficiency are invariant across domains is false. If we add a third or more source domains, it is easy to see that we will have three or more equations but still only two unknowns, and the system is thus over-identified.

Once in possession of PS_{10} and PS_{01}, we can transport the causal effect to the target population Π^* by appealing again to the law of total probability,

$$P_{11}^* = (1 - \mathrm{PS}_{10}) \times P_{01}^* + \mathrm{PS}_{01} \times P_{00}^* \tag{7.150}$$

$$= \frac{P_{11}^a P_{00}^b - P_{11}^b P_{00}^a}{P_{01}^a P_{00}^b - P_{01}^b P_{00}^a} \times P_{01}^* + \frac{P_{11}^b P_{01}^a - P_{11}^a P_{01}^b}{P_{01}^a P_{00}^b - P_{01}^b P_{00}^a} \times P_{00}^* \tag{7.151}$$

Finally, we note that all probabilities of causation, as discussed in [137], are also identifiable in this setting. First, consider the *probability of necessity and sufficiency*, $\mathrm{PNS} = P(Y_1 =$

$1, Y_0 = 0)$ for Π^a. Using the chain rule, PNS can be written as,

$$\begin{aligned} P^a(Y_1 = 1, Y_0 = 0) &= P^a(Y_1 = 1 \mid Y_0 = 0) P^a(Y_0 = 0) && (7.152) \\ &= \mathrm{PS}_{01} \times P^a(Y_0 = 0) && (7.153) \end{aligned}$$

Note PS_{01} was already identified, and $P^a(Y_0 = 0)$ is given by the trial data in Π^a, thus rendering PNS^a identifiable. Similar reasoning holds for Π^b.

For the *probability of necessity*, define $\mathrm{PN}_{01} := P(Y_0 = 0 \mid Y_1 = 1)$. Due to the randomization of X, PN_{01} coincides with Tian and Pear's probability of necessity *during the trial* (not the observational PN), by the same argument we provide for PS in the main text. The final step is to note that,

$$P^a(Y_0 = 0 \mid Y_1 = 1) = \frac{P^a(Y_0 = 0, Y_1 = 1)}{P^a(Y_1 = 1)} = \frac{\mathrm{PNS}^a}{P^a(Y_1 = 1)}$$

The numerator is simply the PNS, which we have already identified, and the denominator is given by the trial data in Π^a. Again, analogous argument can be given for Π^b.

7.4.3 Modeling functional constraints

To illustrate the usefulness of explicitly modeling functional constraints in a structural framework, we apply the same modeling strategy of the paper in an example described in [78]:

> Consider a team of investigators who are interested in the effect of antibiotic treatment on mortality in patients with a specific bacterial infection (...) the investigators believe that the response to this antibiotic is completely determined by an unmeasured bacterial gene, such that only those who are infected with a bacterial strain with this gene respond to treatment. The prevalence of this bacterial gene is equal between populations, because the populations share the same bacterial ecosystem (...) if the investigators further believe that the gene for susceptibility reduces the mortality in the presence of antibiotics, but has no effect in the absence of antibiotics, they will conclude that G may be equal between populations.

Here the conclusion that G may be equal between populations is equivalent to claiming $Y_1 \perp\!\!\!\perp S \mid Y_0$. But is the description above sufficient for substantiating this claim? Figure 7.4 shows two models compatible with the description, yet leading to two opposite conclusions.

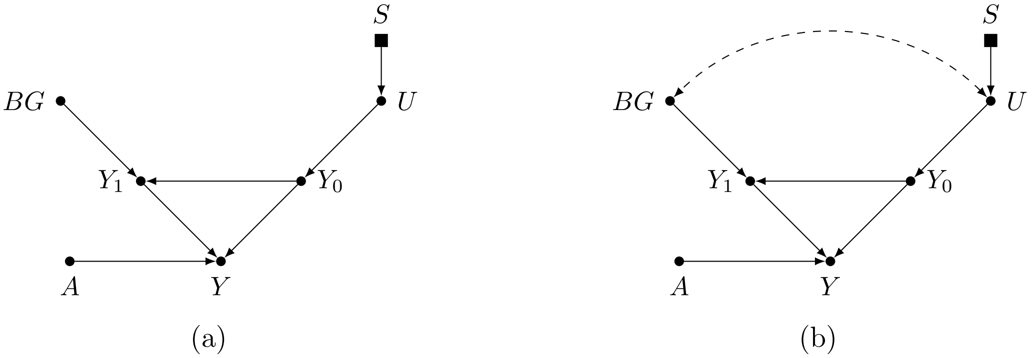


Figure 7.4: Two selection diagrams compatible with the verbal description of [78, page 11]. Yet, model (a) implies $Y_1 \perp\!\!\!\perp S \mid Y_0$, and model (b) implies the opposite; conditioning on Y_0 opens the colliding path $S \to U \leftrightarrow BG \to Y_1$.

Let the variable A represent the binary treatment (antibiotic), Y represent the binary outcome (mortality), BG stand for the presence or absence of the "bacterial gene" and finally let U be a binary variable that summarizes all other factors that may cause death ($Y = 1$). The description of the problem suggests the functional specification,

$$Y = U \wedge (\neg A \vee \neg BG) \tag{7.154}$$

showing the antibiotics and the bacterial gene both helping to *reduce* mortality ($\neg$ denotes the logical "not"). Equation 7.154 entails the potential outcomes $Y_0 = U$ and $Y_1 = U \wedge (\neg BG) = Y_0 \wedge (\neg BG)$, which are explicitly shown in both diagrams as dictated by the functional specification. Moreover, in both models the prevalence of the bacterial gene BG is equal between populations (i.e., $BG \perp\!\!\!\perp S$). In the model of Figure 7.4a, as in our previous analysis, we indeed conclude that $Y_1 \perp\!\!\!\perp S \mid Y_0$, and that $P^*(Y_1)$ is transportable. However, in the model of Figure 7.4b, there is an unmeasured confounder between BG and U.[3] Conditioning on Y_0 (a child of a collider) opens the colliding path $S \to U \leftrightarrow BG \to Y_1$, thus not licensing

[3] This could arise, for instance, as a result of population stratification.

the independence $Y_1 \perp\!\!\!\perp S \mid Y_0$.

7.4.4 Bayesian estimation

7.4.4.1 Multiple source domains

In this section we show how to extend the probabilistic graphical model of Section 5.4 to two or more sources. Let us start with two source populations Π^a and Π^b, and one target domain Π^*. The observed data is now $\mathcal{D} = \{n_0^a, n_1^a, n_0^*, n_0^b, n_1^b\}$, all with binomial distributions:

$$n_0^a \sim \text{Binomial}(N_0^a, P_{01}^a) \tag{7.155}$$

$$n_1^a \sim \text{Binomial}(N_1^a, P_{11}^a) \tag{7.156}$$

$$n_0^* \sim \text{Binomial}(N_0^*, P_{01}^*) \tag{7.157}$$

$$n_0^b \sim \text{Binomial}(N_0^b, P_{01}^b) \tag{7.158}$$

$$n_1^b \sim \text{Binomial}(N_1^b, P_{11}^b) \tag{7.159}$$

We also have the following deterministic relationships for P_{11}^a, P_{11}^b and P_{11}^*:

$$P_{11}^a = (1 - \text{PS}_{10})P_{01}^a + \text{PS}_{01}(1 - P_{01}^a) \tag{7.160}$$

$$P_{11}^b = (1 - \text{PS}_{10})P_{01}^b + \text{PS}_{01}(1 - P_{01}^b) \tag{7.161}$$

$$P_{11}^* = (1 - \text{PS}_{10})P_{01}^* + \text{PS}_{01}(1 - P_{01}^*) \tag{7.162}$$

The probabilistic graphical model for this case is shown in Figure 7.5.

Thus, one needs to place priors on the parent nodes only, and then perform inference as before. The extension to more than two populations follows the same logic. It is worth noting that, as we have seen in Section 5.3, with two or more source populations the model entails testable implications. Therefore, we advise researchers to check whether the data is compatible with the model [68].

Finally, similarly to the discussion in Section 5.4, a simpler modeling alternative here is

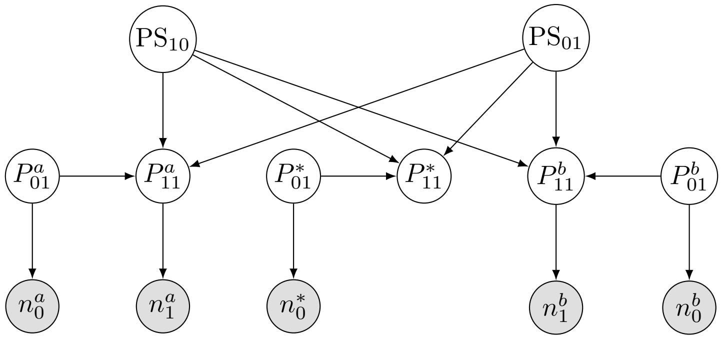


Figure 7.5: Probabilistic graphical model with two source populations Π^a, Π^b and one target population Π^*. Gray nodes (n_0^a, n_1^a, n_0^*, n_0^b, n_1^b) denote observed variables. White notes denote latent parameters (P_{01}^a, P_{11}^a, PS_{10}, PS_{01}, P_{11}^*, P_{01}^*, P_{11}^b, P_{01}^b). Note that P_{11}^a, P_{11}^* and P_{11}^b share the parameters PS_{10} and PS_{01}, which are invariant across populations.

to place priors only on the parameters of the observed data directly, and make inferences using the posterior of the functionals of the observed data that identify the target quantities.

7.4.4.2 Replication code

Here we provide `R` code to replicate the estimation examples using JAGS [115] and the package `rjags` [114].

```
# Set up -------------------------------------------------------


## Cleans workspace
rm(list = ls())
## Loads necessary R packages
library(rjags)


## JAGS models


model_one_source <-
  "model{
```

```
  # Likelihood
  n0 ~ dbinom(p01, N0)
  n1 ~ dbinom(p11, N1)
  n0s ~ dbinom(p01s, N0s)

  # Priors
  PS10 ~ dbeta(1,1)
  PS01 ~ dbeta(1,1)
  p01 ~ dbeta(1, 1)
  p01s ~ dbeta(1, 1)

  # Computed quantities
  p11  <- (1-PS10)*p01  + PS01*(1-p01)
  p11s <- (1-PS10)*p01s + PS01*(1-p01s)
  rd   <- p11s - p01s
  rr   <- p11s/p01s

  # bounds
  PS01_l <- max(0, (p11-p01)/(1-p01))
  PS01_u <- min(p11/(1-p01), 1)
  p11_1 <- (1-p01s/p01)*PS01_l + (p01s/p01)*p11
  p11_2 <- (1-p01s/p01)*PS01_u + (p01s/p01)*p11
  p11_l <- min(p11_1, p11_2)
  p11_u <- max(p11_1, p11_2)
  rd_l <- p11_l - p01s
  rr_l <- p11_l/p01s
}"

model_one_source_monotonic <-
```

```
  "model{

  # Likelihood
  n0 ~ dbinom(p01, N0)
  n1 ~ dbinom(p11, N1)
  n0s ~ dbinom(p01s, N0s)

  # Priors
  PS10 <- 0
  PS01 ~ dbeta(1,1)
  p01 ~ dbeta(1, 1)
  p01s ~ dbeta(1, 1)

  # Computed quantities
  p11  <- (1-PS10)*p01  + PS01*(1-p01)
  p11s <- (1-PS10)*p01s + PS01*(1-p01s)
  rd   <- p11s - p01s
  rr   <- p11s/p01s
}"

model_two_sources <- "model{

  # Likelihood
  n0a ~ dbinom(p01a, N0a)
  n0b ~ dbinom(p01b, N0b)
  n0c ~ dbinom(p01c, N0c)
  n1a ~ dbinom(p11a, N1a)
  n1b ~ dbinom(p11b, N1b)
  # Priors
  p01a ~ dbeta(1, 1)
  p01b ~ dbeta(1, 1)
  p01c ~ dbeta(1, 1)
  PS10 ~ dbeta(1, 1)
  PS01 ~ dbeta(1, 1)

  # Computed quantities
  p11a <- (1-PS10)*p01a + PS01*(1-p01a)
  p11b <- (1-PS10)*p01b + PS01*(1-p01b)
  p11c <- (1-PS10)*p01c + PS01*(1-p01c)
  rra <- (p11a)/(p01a)
  rrb <- (p11b)/(p01b)
  rrc <- (p11c)/(p01c)
}"

# Simulated data example ---------------------------------------
loop_n <- c(1e2, 1e3, 1e4)

### Without monotonicity
par(mfrow = c(1, 3))

for(n in loop_n){
  # creates data
  data <- list(
    N0  = n,
    n0  = sum(rbinom(n, 1, prob = 0.01)),
    N1  = n,
    n1  = sum(rbinom(n, 1, prob = 0.175)),
    N0s = n,
    n0s = sum(rbinom(n, 1, prob = 0.05))
  )

  # posterior samples
  model   <- jags.model(textConnection(model_one_source),
                      data = data)

  samples <- coda.samples(model = model,
                       variable.names = c("p01","p01s", "p11","p11s"),
                       n.iter = 100000)

  samp.data <- as.data.frame(samples[[1]])

  hist(samp.data$p11s,
       main = "",
       xlim = c(0, .4),
       yaxt = "n",
       xaxt = "n",
       xlab = paste0("n = ", n),
       ylab = "",
       col = "gray")

  labs <- round(quantile(data$p11s, c(0.025, 0.975)), 2)
  axis(side = 1, at = c(0, labs, .4))
}

### With monotonicity
par(mfrow = c(1, 3))
```

```
for(n in loop_n){
  data <- list(
    N0  = n,
    n0  = sum(rbinom(n, 1, prob = 0.01)),
    N1  = n,
    n1  = sum(rbinom(n, 1, prob = 0.175)),
    N0s = n,
    n0s = sum(rbinom(n, 1, prob = 0.05))
    )

  # posterior samples
  model   <- jags.model(textConnection(model_one_source_monotonic),
                        data = data)

  samples <- coda.samples(model = model,
                          variable.names = c("p01","p01s", "p11","p11s"),
                          n.iter = 100000)

  samp.data <- as.data.frame(samples[[1]])

  hist(samp.data$p11s,
       main = "",
       xlim = c(0, .4),
       yaxt = "n",
       xaxt = "n",
       xlab = paste0("n = ", n),
       ylab = "",
       col = "gray")
  labs <- round(quantile(data$p11s, c(0.025, 0.975)), 2)
  axis(side = 1, at = c(0, labs, .4))
}

# Vitamin A example --------------------------------------------
### Vitamin A data
### Aceh study
Aceh    <- data.frame(N0 = 12209,
                      n0 = 12079,
                      N1 = 12991,
                      n1 = 12890)

### West Java study
West.Java     <- data.frame(N0 = 5445,
                            n0 = 5195,
                            N1 = 5775,
                            n1 = 5589)

### Sarlahi Study
Sarlahi <- data.frame(N0 = 14143,
                      n0 = 13933,
                      N1 = 14487,
                      n1 = 14335)

## Transporting: Aceh -> West Java
### Data
data <- list(N0  = Aceh$N0,
             n0  = Aceh$n0,
             N1  = Aceh$N1,
             n1  = Aceh$n1,
             N0s = West.Java$N0,
             n0s = West.Java$n0)

### Posterior samples bounds
model.bounds    <- jags.model(textConnection(model_one_source),
                              data = data,  n.chains = 4, n.adapt = 1e3)


## burn-in
update(model.bounds, n.iter = 1e4)


## samples
samp.bounds    <- coda.samples(model.bounds,
                               variable.names = c("p01","p01s", "p11",
                                                  "PS01", "PS10", "p11s",
                                                  "rd", "rr",
                                                  "PS01_l", "PS01_u",
                                                  "p11_l", "p11_u",
                                                  "rd_l", "rr_l"),
                               n.iter = 100000)
summary(samp.bounds)


## extract data.frame
sim.bounds    <- do.call("rbind", samp.bounds)
sim.bounds    <- as.data.frame(sim.bounds)


### Posterior samples monotonic
model.monotonic <- jags.model(textConnection(model_one_source_monotonic),
                              data = data,  n.chains = 4, n.adapt = 1e3)
## burn-in
update(model.monotonic, n.iter = 1e4)
## samples
samp.monotonic <- coda.samples(model.monotonic,
                               variable.names = c("p01","p01s", "p11",
                                                  "PS01", "PS10", "p11s",
                                                  "rd", "rr"),
                               n.iter = 100000)
summary(samp.monotonic)

## extract data.frame
sim.monotonic   <- do.call("rbind", samp.monotonic)
sim.monotonic   <- as.data.frame(sim.monotonic)

## plot
par(mfrow = c(1, 2))
lims <- c(0.94,1)
mark <- West.Java$n1/West.Java$N1

hist(sim.bounds$p11s,
     breaks =  50,
     main = "",
     xlim = lims,
     yaxt = "n",
     xlab = "Flat priors",
     ylab = "",
     col = "gray")
abline(v = mark, col = "red", lty = 2, lwd = 2)
```

```
hist(sim.monotonic$p11s,
     breaks =  50,
     main = "",
     xlim = lims,
     yaxt = "n",
     xlab = "Assuming monotonicity",
     ylab = "",
     col = "gray")
abline(v = mark, col = "red", lty = 2, lwd = 2)

## Transporting: Aceh + West Java -> Sarlahi
### Data
data2 <-  list(N0a = Aceh$N0,
               n0a = Aceh$n0,
               N1a = Aceh$N1,
               n1a = Aceh$n1,
               N0b = West.Java$N0,
               n0b = West.Java$n0,
               N1b = West.Java$N1,
               n1b = West.Java$n1,
               N0c = Sarlahi$N0,
               n0c = Sarlahi$n0)

### Posterior samples two sources
model2    <- jags.model(textConnection(model_two_sources),
                        data = data2,  n.chains = 4, n.adapt = 1e3)

## burn in
```

```
update(model2, n.iter = 1e4)
## samples
samp2 <- coda.samples(model2,
                      variable.names = c("p01a","p01b","p01c",
                                         "p11a", "p11b","p11c",
                                         "PS01", "PS10",
                                         "rra", "rrb", "rrc"),
                      n.iter = 100000)
summary(samp2)

## extract data.frame
sim2 <- as.data.frame(samp2[[1]])

### Plot
par(mfrow = c(1, 3))
mark <- Sarlahi$n1/Sarlahi$N1
hist(sim2$PS01, xlim = c(0,1), breaks = 50,
     yaxt = "n", col = "gray", main = "", xlab = "PS01", ylab = "")
hist(sim2$PS10, xlim = c(0, 0.1),
     yaxt = "n", col = "gray", main = "", xlab = "PS10", ylab = "")
hist(sim2$p11c,
     yaxt = "n", col = "gray", main = "", xlab = "P11*", ylab = "")
abline(v = mark, col = "red", lty = 2, lwd = 2)
```

Bibliography

[1] Joseph G Altonji, Todd E Elder, and Christopher R Taber. An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *Journal of Human resources*, 40(4):791–821, 2005.

[2] Theodore W Anderson and Herman Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63, 1949.

[3] Isaiah Andrews, James H Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753, 2019.

[4] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[5] Joshua D Angrist and Alan B Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85, 2001.

[6] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2008.

[7] Joshua D Angrist and Jörn-Steffen Pischke. *Mastering 'metrics: The path from cause to effect*. Princeton University Press, 2014.

[8] Joshua D Angrist and Jörn-Steffen Pischke. Undergraduate econometrics instruction: Through our classes, darkly. Technical report, National Bureau of Economic Research, 2017.

[9] Onyebuchi A Arah. Bias analysis for uncontrolled confounding in the health sciences. *Annual review of public health*, 38:23–38, 2017.

[10] Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014.

[11] Ioana Baldini, Clark W. Barrett, Antonio Chella, Carlos Cinelli, David Gamez, Leilani H. Gilpin, Knut Hinkelmann, Dylan Holmes, Takashi Kido, Murat Kocaoglu, William F. Lawless, Alessio Lomuscio, Jamie C. Macbeth, Andreas Martin, Ranjeev Mittu, Evan Patterson, Donald Sofge, Prasad Tadepalli, Keiki Takadama, and Shomir Wilson. Reports of the AAAI 2019 spring symposium series. *AI Mag.*, 40(3):59–66, 2019.

[12] Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In Ramón López de Mántaras and David Poole, editors, *UAI '94: Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence, Seattle, Washington, USA, July 29-31, 1994*, pages 46–54. Morgan Kaufmann, 1994.

[13] Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

[14] M Bardet. Algorithms seminar. *On the complexity of a Groebner basis algorithm*, pages 2002–2004, 2005.

[15] Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. In Nando de Freitas and Kevin P. Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pages 113–120. AUAI Press, 2012.

[16] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[17] Matthew Blackwell. A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, 22(2):169–182, 2013.

[18] John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endoge-

nous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450, 1995.

[19] Roger J Bowden and Darrell A Turkington. *Instrumental variables*, volume 8. Cambridge University Press, 1990.

[20] Carlos Brito and Judea Pearl. Generalized instrumental variables. In Adnan Darwiche and Nir Friedman, editors, *UAI '02, Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence, University of Alberta, Edmonton, Alberta, Canada, August 1-4, 2002*, pages 85–93. Morgan Kaufmann, 2002.

[21] Babette A Brumback, Miguel A Hernán, Sebastien JPA Haneuse, and James M Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in medicine*, 23(5):749–767, 2004.

[22] Stephen Burgess and Simon G Thompson. *Mendelian randomization: methods for using genetic variants in causal estimation.* CRC Press, 2015.

[23] David Card. Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research, 1993.

[24] David Card. The causal effect of education on earnings. In *Handbook of labor economics*, volume 3, pages 1801–1863. Elsevier, 1999.

[25] Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016.

[26] Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016.

[27] Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. treatsens: A package to assess sensitivity of causal analyses to unmeasured confounding, 2016.

[28] Bryant Chen, Daniel Kumor, and Elias Bareinboim. Identification and model testing in linear structural equation models using auxiliary variables. In *International Conference on Machine Learning*, pages 757–766, 2017.

[29] Bryant Chen and Judea Pearl. Exogeneity and robustness. Technical report, UCLA Cognitive Systems Laboratory, 2015.

[30] Bryant Chen, Judea Pearl, and Elias Bareinboim. Incorporating knowledge into structural equation models using auxiliary variables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*, pages 3577–3583, 2016.

[31] Carlos Cinelli, Jeremy Ferwerda, and Chad Hazlett. sensemakr: Sensitivity analysis tools for OLS in R and Stata. *SSRN Electronic Journal*, Abstract ID 3588978, 2020.

[32] Carlos Cinelli, Jeremy Ferwerda, and Chad Hazlett. Sensemakr: Stata module to provide sensitivity tools for ols. *Boston College Department of Economics: Statistical Software Components*, 2020.

[33] Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *SSRN Electronic Journal*, Abstract ID 3689437, 2020.

[34] Carlos Cinelli and Chad Hazlett. *sensemakr: Sensitivity Analysis Tools for OLS*, 2019. R package version 0.1.2.

[35] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

[36] Carlos Cinelli and Chad Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Working Paper*, 2020.

[37] Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, and Elias Bareinboim. Sensitivity analysis of linear structural causal models. In Kamalika Chaudhuri and Ruslan

Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1252–1261. PMLR, 2019.

[38] Carlos Cinelli, Nathan LaPierre, Brian Hill, Sriram Sankararaman, and Eleazar Eskin. Robust mendelian randomization in the presence of residual population stratification, batch effects and horizontal pleiotropy. *bioRxiv*, 2020.

[39] Carlos Cinelli and Judea Pearl. On the utility of causal diagrams in modeling attrition: a practical example. *Epidemiology*, 29(6):e50–e51, 2018.

[40] Carlos Cinelli and Judea Pearl. Generalizing experimental results by leveraging knowledge of mechanisms. *European Journal of Epidemiology*, 36:149 – 164, 2021.

[41] Timothy G Conley, Christian B Hansen, and Peter E Rossi. Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272, 2012.

[42] Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *journal of National Cancer Institute*, 22(1), 1959.

[43] David Cox, John Little, and Donal O'shea. *Ideals, Varieties, and Algorithms*, volume 3. Springer, 1992.

[44] Issa J Dahabreh, Lucia C Petito, Sarah E Robertson, Miguel A Hernán, and Jon A Steingrimsson. Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology*, 31(3):334–344, 2020.

[45] Angus S Deaton. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Technical report, National bureau of economic research, 2009.

[46] Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, 16(4):309–330, 2007.

[47] Peng Ding and Luke W Miratrix. To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference*, 3(1):41–57, 2015.

[48] Peng Ding and Tyler J VanderWeele. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368, 2016.

[49] Thomas A DiPrete and Markus Gangl. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological methodology*, 34(1):271–310, 2004.

[50] Vincent Dorie, Masataka Harada, Nicole Bohme Carnegie, and Jennifer L Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470, 2016.

[51] Thad Dunning. *Natural experiments in the social sciences: a design-based approach*. Cambridge University Press, 2012.

[52] Edgar C Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185, 1954.

[53] Ronald A Fisher. Dangers of cigarette-smoking. *British Medical Journal*, 2(5039):297, 1957.

[54] Ronald A Fisher. Cigarettes, cancer, and statistics. *The Centennial Review of Arts & Science*, 2:151–166, 1958.

[55] Julie Flint and Alex de Waal. *Darfur: a new history of a long war*. Zed Books, 2008.

[56] Victor Fossaluza, Rafael Izbicki, Gustavo Miranda da Silva, and Luís Gustavo Esteves. Coherent hypothesis testing. *The American Statistician*, 71(3):242–248, 2017.

[57] Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, pages 1682–1713, 2012.

[58] Kenneth Frank. Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, 29(2):147–194, 2000.

[59] Kenneth Frank and Kyung-Seok Min. Indices of robustness for sample representation. *Sociological Methodology*, 37(1):349–392, 2007.

[60] Kenneth Frank, Gary Sykes, Dorothea Anagnostopoulos, Marisa Cannata, Linda Chard, Ann Krause, and Raven McCrory. Does NBPTS certification affect the number of colleagues a teacher helps with instructional matters? *Educational Evaluation and Policy Analysis*, 30(1):3–30, 2008.

[61] Kenneth A Frank, Spiro J Maroulis, Minh Q Duong, and Benjamin M Kelcey. What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, 35(4):437–460, 2013.

[62] Alexander M Franks, Alexander D'Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 115(532):1730–1746, 2020.

[63] Ragnar Frisch and Frederick V Waugh. Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401, 1933.

[64] K Ruben Gabriel. Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, pages 224–250, 1969.

[65] Trevor Gallen. Broken instruments. *SSRN Electronic Journal*, Abstract ID 3671850, 2020.

[66] Luis Garcia, Sarah Spielvogel, and Seth Sullivant. Identifying causal effects with computer algebra. In Peter Grünwald and Peter Spirtes, editors, *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, pages 193–200. AUAI Press, 2010.

[67] Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

[68] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[69] Robert B Giffin, Yeonwoo Lebovitz, Rebecca A English, et al. *Transforming clinical research in the United States: challenges and opportunities: workshop summary*. National Academies Press, 2010.

[70] Paul Gustafson. *Bayesian inference for partially identified models: Exploring the limits of limited data*. CRC Press, 2015.

[71] Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sate to patt: combining experimental with observational studies to estimate population treatment effects. *Journal of Royal Statistical Society Series A (Statistics in Society)*, 10:1111, 2015.

[72] Chad Hazlett. Angry or weary? how violence impacts attitudes toward peace among darfurian refugees. *Journal of Conflict Resolution*, 64(5):844–870, 2020.

[73] James J Heckman and Sergio Urzua. Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics*, 156(1):27–37, 2010.

[74] Miguel A Hernán and James M Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

[75] Guanglei Hong, Xu Qin, and Fan Yang. Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics*, 43(1):32–56, 2018.

[76] Carrie A Hosman, Ben B Hansen, and Paul W Holland. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics*, 4(2):849–870, 2010.

[77] Anders Huitfeldt. Effect heterogeneity and external validity in medicine. Available in: https://www.lesswrong.com/posts/wwbrvumMWhDfeo652/, 2019.

[78] Anders Huitfeldt, Andrew Goldstein, and Sonja A Swanson. The choice of effect measure for binary outcomes: Introducing counterfactual outcome state transition parameters. *Epidemiologic methods*, 7(1):20160014, 2018.

[79] Anders Huitfeldt, Sonja A Swanson, Mats J Stensrud, and Etsuji Suzuki. Effect heterogeneity and variable selection for standardizing causal effects to a target population. *European journal of epidemiology*, 34(12):1119–1129, 2019.

[80] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, 25(1):51–71, 2010.

[81] Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review*, 93(2):126–132, 2003.

[82] Guido W Imbens. Instrumental variables: An econometrician's perspective. Technical report, National Bureau of Economic Research, 2014.

[83] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

[84] Pamela Jakiela and Owen Ozier. Gendered language. *Policy Research Working Paper, World Bank*, 2018.

[85] Yang Jiang, Hyunseung Kang, and Dylan S Small. ivmodel: An r package for inference and sensitivity analysis of instrumental variables models with one endogenous variable. *R package vignette*, 2018.

[86] Marshall M Joffe, Wei Peter Yang, and Harold I Feldman. Selective ignorability assumptions in causal inference. *The International Journal of Biostatistics*, 6(2), 2010.

[87] Désiré Kédagni and Ismael Mourifié. Generalized instrumental inequalities: testing the instrumental variable independence assumption. *Biometrika*, 107(3), 2020.

[88] William Kruskal and Ruth Majors. Concepts of relative importance in recent scientific literature. *The American Statistician*, 43(1):2–6, 1989.

[89] Daniel Kumor, Carlos Cinelli, and Elias Bareinboim. Efficient identification in linear structural causal models with auxiliary cutsets. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5501–5510. PMLR, 2020.

[90] Edward E Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43, 1983.

[91] Edward E Leamer. S-values and bayesian weighted all-subsets regressions. *European Economic Review*, 81:15–31, 2016.

[92] Edward E Leamer. S-values: Conventional context-minimal measures of the sturdiness of regression coefficients. *Journal of Econometrics*, 193(1):147 – 161, 2016.

[93] Michael C Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963.

[94] Michael C Lovell. A simple proof of the FWL theorem. *The Journal of Economic Education*, 39(1):88–91, 2008.

[95] Yi Lu, Daniel O Scharfstein, Maria M Brooks, Kevin Quach, and Edward H Kennedy. Causal inference for comprehensive cohort studies. *arXiv preprint arXiv:1910.03531*, 2019.

[96] Matthew A Masten and Alexandre Poirier. Identification of treatment effects under conditional partial independence. *Econometrica*, 86(1):317–351, 2018.

[97] Robert Mauro. Understanding love (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, 108(2):314, 1990.

[98] Halvor Mehlum. The polar confidence curve for a ratio. *Econometric Reviews*, 39(3):234–243, 2020.

[99] Jonathan Mellon. Rain, rain, go away: 176 potential exclusion-restriction violations for studies using weather as an instrumental variable. *SSRN Electronic Journal*, Abstract ID 3715610, 2020.

[100] Joel A Middleton, Marc A Scott, Ronli Diakow, and Jennifer L Hill. Bias amplification and bias unmasking. *Political Analysis*, 24(3):307–323, 2016.

[101] Scott Mueller and Judea Pearl. Which patients are in greater need: A counterfactual analysis with reflections on COVID-19. *Causal Analysis in Theory and Practice*, 2020. Available in: https://ucla.in/39Ey8sU.

[102] Permeisih D Muhilal, Yanyan R Idjradinata, and Karyadi D Muherdiyantiningsih. Vitamin a-fortified monosodium glutamate and health, growth, and survival of children: a controlled field trial. *Am J Clin Nutr*, 48(5):1271–1276, 1988.

[103] Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *NBER working paper*, 2014.

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

[105] Alexandre G Patriota. A classical measure of evidence for general null hypotheses. *Fuzzy Sets and Systems*, 233:74–88, 2013.

[106] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[107] Judea Pearl. On the testability of causal models with latent and instrumental variables. In Philippe Besnard and Steve Hanks, editors, *UAI '95: Proceedings of the Eleventh*

Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995, pages 435–443. Morgan Kaufmann, 1995.

[108] Judea Pearl. Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, 121(1-2):93–149, 1999.

[109] Judea Pearl. *Causality.* Cambridge University Press, 2009.

[110] Judea Pearl. Invited commentary: understanding bias amplification. *American journal of epidemiology*, 174(11):1223–1227, 2011.

[111] Judea Pearl. Causes of effects and effects of causes. *Sociological Methods & Research*, 44(1):149–164, 2015.

[112] Judea Pearl. Sufficient causes: On oxygen, matches, and fires. *Journal of Causal Inference*, 7(2), 2019.

[113] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

[114] Martyn Plummer. rjags: Bayesian graphical models using MCMC. *R package version*, 4(6), 2016.

[115] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria., 2003.

[116] Thomas S Richardson, Robin J Evans, and James M Robins. Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610, 2011.

[117] James M Robins. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pages 113–159, 1989.

[118] James M Robins. Association, causation, and marginal structural models. *Synthese*, 121(1):151–179, 1999.

[119] Paul R Rosenbaum. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147(5):656–666, 1984.

[120] Paul R Rosenbaum. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.

[121] Paul R Rosenbaum. Sensitivity analysis in observational studies. In *Encyclopedia of statistics in behavioral science*, volume 4, pages 1809, 1814. John Wiley & Sons Ltd, 2005.

[122] Paul R Rosenbaum. *Design of observational studies.* Springer Series in Statistics, 2010.

[123] Paul R Rosenbaum. *Observation and experiment: an introduction to causal inference.* Harvard University Press, 2017.

[124] Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218, 1983.

[125] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[126] Mark J Schervish. P values: what they are and what they are not. *The American Statistician*, 50(3):203–206, 1996.

[127] Ilya Shpitser, Tyler VanderWeele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*, 2012.

[128] Ricardo Silva and Robin Evans. Causal inference through a witness protection program. *The Journal of Machine Learning Research*, 17(1):1949–2001, 2016.

[129] Dylan S Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007.

[130] Dylan S Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007.

[131] Dylan S Small and Paul R Rosenbaum. War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, 103(483):924–933, 2008.

[132] Alfred Sommer, Edi Djunaedi, A A Loeden, Ignatius Tarwotjo, Keith P West Jr, Robert Tilden, Lisa Mele, Aceh Study Group, et al. Impact of vitamin a supplementation on childhood mortality: a randomised controlled community trial. *The Lancet*, 327(8491):1169–1173, 1986.

[133] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[134] Peter M Steiner and Yongnam Kim. The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of causal inference*, 4(2):20160009, 2016.

[135] Sonja A Swanson, Miguel A Hernán, Matthew Miller, James M Robins, and Thomas S Richardson. 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, 2018.

[136] The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 8.5)*, 2018. https://www.sagemath.org.

[137] Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

[138] Jin Tian and Judea Pearl. On the identification of causal effects. Technical Report R-290, Cognitive Systems Laboratory, UCLA, 2003.

[139] Santtu Tikka, Antti Hyttinen, and Juha Karvanen. Identifying causal effects via context-specific independence relations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2800–2810, 2019.

[140] Tyler J VanderWeele. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass.)*, 21(4):540, 2010.

[141] Tyler J Vanderweele and Onyebuchi A Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, 22(1):42–52, January 2011.

[142] Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

[143] Xuran Wang, Yang Jiang, Nancy R Zhang, and Dylan S Small. Sensitivity analysis and power for instrumental variable studies. *Biometrics*, 74(4):1150–1160, 2018.

[144] Keith P West Jr, Joanne Katz, Steven Charles LeClerq, EK Pradhan, James M Tielsch, Alfred Sommer, RP Pokhrel, SK Khatry, SR Shrestha, and MR Pandey. Efficacy of vitamin a in reducing preschool child mortality in nepal. *The Lancet*, 338(8759):67–71, 1991.

[145] Philip G Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

[146] Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

[147] Chi Zhang, Carlos Cinelli, Bryant Chen, and Judea Pearl. Exploiting equality constraints in causal inference. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1630–1638. PMLR, 2021.