Causal Mediation in Natural Experiments

Senan Hogan-Hennessy*
Economics Department, Cornell University†

First draft: 12 February 2025
This version: 5 May 2025

Work in Progress, newest version available here.

Abstract

Natural experiments are a cornerstone of applied economics, providing settings for estimating causal effects with a compelling argument for treatment ignorability. Applied researchers often investigate mechanisms behind treatment effects by controlling for a mediator of interest, alluding to Causal Mediation (CM) methods for estimating direct and indirect effects (CM effects). This approach to investigating mechanisms unintentionally assumes the mediator is quasi-randomly assigned — in addition to the quasi-random assignment of the initial treatment. Individuals' choice to take (or refuse) a mediator based on costs and benefits is inconsistent with mediator ignorability, suggesting in-practice estimates of CM effects are biased in natural experiment settings. I solve for explicit bias terms when the mediator is not ignorable, imitating classical selection bias for average causal effects. I consider an alternative approach to credibly estimate CM effects, when selection-into=mediator is driven by unobserved costs and benefits. The approach uses a control function adjustment, relying on mediator take-up cost as an instrument. Simulations confirm that this method corrects for selection bias in conventional CM estimates, providing both parametric and semi-parametric methods. This approach gives applied researchers an alternative method to estimate CM effects when they can only establish a credible argument for quasi-random assignment of the initial treatment, and not a mediator, as is common in natural experiments.

Keywords: Direct/indirect effects, quasi-experiment, selection, control function.
JEL Codes: C21, C31.

*This work was completed while receiving a Sage Research Fellowship from Cornell University. For helpful comments I thank Lexin Cai, Neil Cholli, Hyewon Kim, Jiwoo Kim, Bart de Koning, Lukáš Lafférs, Jiwon Lee, Yiqi Liu, Douglas Miller, Zhuan Pei, Brenda Prallon, Evan Riehl, and Yiwei Sun. Some preliminary results previously circulated in an earlier version of the working paper "The Direct and Indirect Effects of Genetics and Education." I thank seminar participants at Cornell University (2025) for helpful discussion. Any comments or suggestions may be sent to me at seh325@cornell.edu, or raised as an issue on the Github project.

†Address: Uris Hall #447, Economics Department, Cornell University NY 14853 USA.

Economists use natural experiments to credibly answer social questions, when an experiment was infeasible. For example, does gaining access to health insurance causally benefit health outcomes (Finkelstein, Taubman, Wright, Bernstein, Gruber, Newhouse, Allen, Baicker & Group 2012)? Natural experiments are settings which give answers to these questions, but give no indication of how these effects came about. Causal Mediation (CM) aims to estimate the mechanisms behind causal effects, by estimating how much of the treatment effect operates through a proposed mediator. For example, how much of the health gain from new access to health insurance comes from individuals choosing to visit hospitals more frequently, when previously they would have avoided doing so? This study of mechanisms behind causal effects broadens the economic understanding of social setting studied with natural experiments. This paper shows that the conventional approach to estimating CM effects is inappropriate in a natural experiment setting, provides a theoretical framework for how bias operates, and develops an approach to correctly estimate CM effects under alternative assumptions.

This paper starts by answering the following question: what does a selection-on-observables approach to CM actually estimate when a mediator is not quasi-randomly assigned? Estimates for the average direct and indirect effects are contaminated by bias terms — selection bias plus group difference terms. I then show how this bias operates in an applied regression framework, with bias coming from a correlated error term. For example, if individuals had been choosing to seek medical care more frequently after when they have health insurance, then mediator ignorability would require a researcher observe and control for everyone's underlying health conditions as part of the CM analysis — even health conditions that are not yet diagnosed. Should a researcher consider running a CM analysis without using another natural experiment to isolate random variation in the mediator (in addition to the one for the original treatment), then this assumption is unlikely to hold true. This means that investigating mechanisms by CM methods will lead to biased inference in natural experiment settings.

I consider an alternative approach to estimating CM effects, adjusting for unobserved selection-into-mediator with a control function adjustment. This solves the identification problem with structural assumptions for selection-into-mediator — mediator monotonicity and selection based on costs and benefits — and requires a valid cost instrument for mediator take-up. While these assumptions are strong, they are plausible in many applied settings. Mediator monotonicity aligns with conventional theories for selection-into-treatment, and is accepted widely in many applications using an instrumental variables research design. Selection based on costs and benefits is central to economic theory, and is the dominant concern for judging empirical designs that use quasi-experimental variation to estimate causal effects. Access to a valid instrument is a strong assumption, though is important to avoid further modelling assumptions; the most compelling example is using variation in mediator take-up cost as an instrument. This approach is not perfect in every setting: the structural assumptions are strong, and are tailored to selection-into-mediator concerns pertinent to economic applications. Indeed, this approach provides no safe harbour for estimating CM effects if these structural assumptions do not hold true.

The most popular approach to CM assumes that the original treatment, and the subsequent mediator, are both ignorable (Imai, Keele & Yamamoto 2010). This approach arose in the statistics literature, and is widely used in social sciences to estimate CM effects in observational studies. Informal investigations of mechanisms in applied economics are fundamentally CM analyses, and so unintentionally import CM's identifying assumptions.

Assuming the mediator is ignorable (i.e., quasi-randomly assigned or satisfies selection-on-observables) conveniently sidesteps the problem of individual choice by assuming that either people made decisions to take/refuse a mediator naïvely, or a researcher has observed everything relevant to this decision. This assumption might be reasonable when studying single-celled organisms in a laboratory — their "decisions" are simple and mechanical. Social scientists, however, study humans who make complex choices based on costs, benefits, and preferences — which are only partially observed by researchers, at best. Assuming a mediator is ignorable in social science contexts is often unrealistic. In practice, the only setting where mediator ignorability becomes credible is when researchers find another natural experiment affecting the mediator — a rare occurrence given how difficult it is to find one source of random variation, let alone two, simultaneously.

The applied economics literature has been hesitant to use explicit CM methods, and began conducting informal mechanism analyses by analysing causal effects when controlling for a proposed mediator (Blackwell, Ma & Opacic 2024). This practice is fundamentally a CM analysis, despite not being named so explicitly, so falls prey to the assumptions of conventional CM analyses just the same. A new strand of the econometric literature has developed estimators for explicit CM analyses under a variety of strategies to avoid relying on unrealistic assumptions. This includes overlapping quasi-experimental research designs (Deuchert, Huber & Schelker 2019, Frölich & Huber 2017, Heckman & Pinto 2015), partial identification (Flores & Flores-Lagunes 2009), or a hypothesis test of full mediation through observed channels (Kwon & Roth 2024) — see Huber (2020) for an overview. The new literature has arisen in implicit acknowledgement that a conventional selection-on-observables approach to CM in applied settings can lead to biased inference, and needs alternative methods for credible inference.

This paper explicitly shows how a conventional approaches to CM can lead to biased inference in natural experiments. I develop a formal framework showing exactly how selection bias contaminates CM estimates when mediator choices are driven by unobserved gains — settings where none of the natural experiment research designs in the previously cited papers apply (i.e., the mediator is not ignorable). This provides a rigorous warning to applied economists against uncritically applying conventional CM methods to investigate mechanisms in natural experiments. Instead, I propose an alternative approach grounded in classic labour economic theory.

I use the Roy (1951) model as a benchmark for judging the Imai et al. (2010) mediator ignorability assumption in a natural experiment setting, and find it unlikely to hold in practice.[1] This motivates a solution to the identification problem inspired by classic labour economic work, which also uses the Roy model as a benchmark (Heckman 1979, Heckman & Honore 1990). I follow the lead of these papers by using a control function to correct for the selection bias in conventional CM analyses.

The control function approach requires a mediator take-up respond only positively to an initial treatment, exploiting the monotonicity–selection model equivalence result in a mediation setting (Vytlacil 2002). Second, it assumes that mediator take-up is motivated by mediator costs and benefits (Florens, Heckman, Meghir & Vytlacil 2008). Last, it requires a valid instrument for mediator take-up, to avoid relying on parametric assumptions on unobserved selection (Heckman & Navarro-Lozano 2004). While these assumptions are strong, they are plausible in many applied settings.

This approach to identifying CM effects (despite selection-into-mediator) imports insights from an instrumental variables setting (Kline & Walters 2019, Heckman & Vytlacil 2005), to account for selection and identifying mediator compliers for CM effects. Doing so is related to using instruments to identifying complier CM effects with two instrumental variables (Frölich & Huber 2017).[2] Using a control function to estimate CM effects builds on the influential Imai et al. (2010) approach, marrying the CM literature with labour economic theory on selection-into-treatment for the first time.

This paper proceeds as follows. Section 1 introduces the formal framework for CM, and develops expressions for bias in CM estimates in natural experiments. Section 2 describes this bias in applied settings with (1) a regression framework, (2) a setting with selection based on costs and benefits. Section 3 shows how a control function can effectively purge

[1] An alternative method to estimate CM effects is ensuring treatment and mediator ignorability holds by a running randomised controlled trial (or suitable quasi-experiment) for both treatment and mediator, at the same time. This set-up has been considered in the literature previously, in theory (Imai, Tingley & Yamamoto 2013, Heckman & Pinto 2015) and in practice (Ludwig, Kling & Mullainathan 2011, Heckman, Pinto & Savelyev 2013).

[2] Indeed, this paper does not improve on selection model/control function methods, instead noting their applicability in this setting. See Frölich & Huber (2017) for the newest development of control function methods with two instruments, and Wooldridge (2015) for a general overview of the approach.

this bias from CM estimates. Section 4 demonstrates how to estimate CM effects with this approach, with either parametric and semi-parametric methods, giving supporting simulation evidence. Section 5 concludes.

1 Direct and Indirect Effects

Causal mediation decomposes causal effects into two channels, through a mediator (indirect effect) and through all other paths (direct effect). To develop notation for direct and indirect effects, write Z_i for an exogenous binary treatment, D_i a binary mediator, and Y_i an outcome for individuals $i = 1, \dots, N$. The outcomes are a sum of their potential outcomes.[3]

$$D_i = Z_i D_i(1) + (1 - Z_i) D_i(0),$$
$$Y_i = Z_i Y_i(1, D_i(1)) + (1 - Z_i) Y_i(0, D_i(0)).$$

Assume Z_i is ignorable.[4]

$$Z_i \perp\!\!\!\perp D_i(z), Y_i(z', d), \text{ for } z, z', d = 0, 1$$

There are only two average effects which are identified (without additional assumptions).

1. The average first-stage refers to the effect of the treatment on mediator, $Z \to D$:

$$\mathbb{E}\left[D_i \,|\, Z_i = 1\right] - \mathbb{E}\left[D_i \,|\, Z_i = 0\right] = \mathbb{E}\left[D_i(1) - D_i(0)\right].$$

It is common in the economics literature to assume that Z influences D in at most one direction, $\Pr\left(D_i(1) \geq D_i(0)\right) = 1$ — monotonicity (Imbens & Angrist 1994). I assume mediator monotonicity (and its conditional variant) holds throughout to simplify notation.

2. The Average Treatment Effect (ATE) refers to the effect of the treatment on outcome, $Z \to Y$, and is also known as the average total effect or intent-to-treat effect in social science settings, or reduced-form effect in the instrumental variables literature:

$$\mathbb{E}\left[Y_i \,|\, Z_i = 1\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0\right] = \mathbb{E}\left[Y_i(1, D_i(1)) - Y_i(0, D_i(0))\right].$$

Z affects outcome Y directly, and indirectly via the $D(Z)$ channel, with no reverse causality. Figure 1 visualises the design, where the direction arrows denote the causal

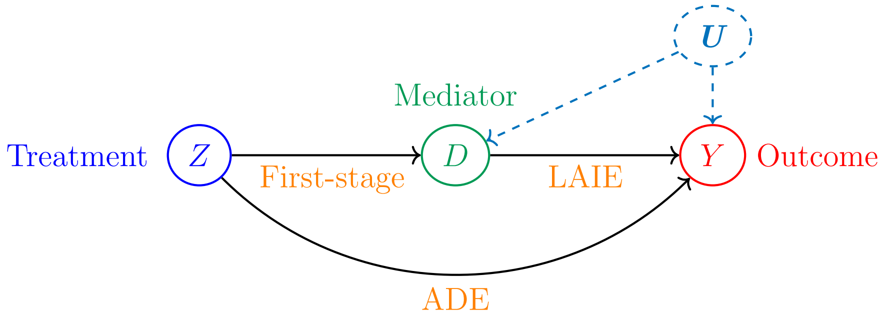
[3] This paper exclusively focuses on the binary case. See Huber, Hsu, Lee & Lettry (2020) for a discussion of CM with continuous treatment and/or mediator, and the assumptions required.

[4] This assumption can hold conditional on covariates. To simplify notation in this section, leave the conditional part unsaid, as it changes no part of the identification framework.

direction. CM aims to decompose the ATE of $Z \to Y$ into these two separate pathways:

$$\text{Average Direct Effect (ADE), } Z \to Y: \quad \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right],$$
$$\text{Average Indirect Effect (AIE), } D(Z) \to Y: \quad \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right].$$

Figure 1: Structural Causal Model for Causal Mediation.



Note: This figure shows the structural causal model behind causal mediation. LAIE refers to the AIE (i.e., effect of the mediator $D \to Y$) local to $D(Z)$ compliers, so that AIE = average first-stage $\times$ LAIE. Unobserved confounder $\boldsymbol{U}$ represents this paper's focus on the case that D is not ignorable, by showing an unobserved confounder. Subsection 2.1 formally defines $\boldsymbol{U}$ in an applied setting.

Estimating the AIE answers the following question: how much of the causal effect $Z \to Y$ goes through the D channel? If a researcher is studying the health gains of health insurance (Finkelstein et al. 2012), and wants to study the role of healthcare usage, the AIE represents how much of the effect comes from using the hospital more often. Estimating the ADE answers the following equation: how much is left over after accounting for the D channel?[5] For the health insurance example, how much of the health insurance effect is a direct effect, other than increased healthcare usage — e.g., long-term effects of lower medical debt, or less worry over health shocks. An instrumental variables approach assumes this direct effect is zero for everyone (the exclusion restriction). CM is a similar, yet distinct, framework attempting to explicitly model the direct effect, and not assuming it is zero.

The ADE and AIE are not separately identified without further assumptions.

1.1 Identifying Causal Mediation (CM) Effects

The conventional approach to estimating direct and indirect effects assumes both Z_i and D_i are ignorable, conditional on a vector of control variables $\boldsymbol{X}_i$.

[5] In a non-parametric setting it is not necessary that ADE + AIE = ATE. See Imai et al. (2010) for this point in full.

Definition 1. *Sequential Ignorability (Imai et al. 2010).*

$$Z_i \perp\!\!\!\perp D_i(z), Y_i(z', d) \mid \boldsymbol{X}_i, \qquad \text{for } z, z', d = 0, 1 \tag{1}$$

$$D_i \perp\!\!\!\perp Y_i(z', d) \mid \boldsymbol{X}_i, Z_i = z', \qquad \text{for } z', d = 0, 1 \tag{2}$$

Sequential ignorability assumes that the initial treatment Z_i is ignorable conditional on $\boldsymbol{X}_i$. It then also assumes that, after Z_i is assigned, that D_i is ignorable conditional on $\boldsymbol{X}, Z$. In addition, a common support condition for both Z_i, D_i (across $\boldsymbol{X}_i$) is necessary. If sequential ignorability, 1(1) and 1(2), holds then the ADE and AIE are identified by two-stage mean differences, after conditioning on $\boldsymbol{X}_i$.[6]

$$\mathbb{E}_{D_i, \boldsymbol{X}_i}\left[\underbrace{\mathbb{E}\left[Y_i \mid Z_i = 1, D_i, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \mid Z_i = 0, D_i, \boldsymbol{X}_i\right]}_{\text{Second-stage regression, } Y_i \text{ on } Z_i \text{ holding } D_i, \boldsymbol{X}_i \text{ constant}}\right] = \underbrace{\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right]}_{\text{Average Direct Effect (ADE)}}$$

$$\mathbb{E}_{Z_i, \boldsymbol{X}_i}\left[\underbrace{\Big(\mathbb{E}\left[D_i \mid Z_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[D_i \mid Z_i = 0, \boldsymbol{X}_i\right]\Big)}_{\text{First-stage regression, } D_i \text{ on } Z_i} \times \underbrace{\Big(\mathbb{E}\left[Y_i \mid Z_i, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \mid Z_i, D_i = 0, \boldsymbol{X}_i\right]\Big)}_{\text{Second-stage regression, } Y_i \text{ on } D_i \text{ holding } Z_i, \boldsymbol{X}_i \text{ constant}}\right]$$

$$= \underbrace{\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right]}_{\text{Average Indirect Effect (AIE)}}$$

I refer to the estimands on the left-hand side as Causal Mediation (CM) estimands. These estimands are typically estimated with linear models, with resulting estimates composed from two-stage Ordinary Least Squares (OLS) estimates (Imai et al. 2010). While this is the most common approach in the applied literature, I do not assume the linear model. Linearity assumptions are unnecessary to my analysis; it suffices to note that heterogeneous treatment effects and non-linear confounding would bias OLS estimates of CM estimands in the same manner that is well documented elsewhere (see e.g., Angrist 1998, Słoczyński 2022). This section focuses on problems that plague CM by selection-on-observables, regardless of estimation method.

1.2 Bias in Causal Mediation (CM) Estimates

Applied researchers often use a natural experiment to study settings where treatment Z_i is ignorable, justifying assumption 1(1). Rarely do they also have access to additional, overlapping natural experiment to isolate random variation D_i to justify assumption 1(2) as

[6] Imai et al. (2010) show a general identification statement; I show identification in terms of two-stage regression, notation for which is more familiar in economics. This reasoning is in line with G-computation reasoning (Robins 1986); Appendix A.1 states the Imai et al. (2010) identification result, and then develops the two-stage regression notation which holds as a consequence of sequential ignorability.

part of the same analysis. One might consider conventional CM methods in such a setting to learn about the mechanisms behind the causal effect $Z \to Y$ under study. This approach leads to biased estimates, and contaminates inference regarding direct and indirect effects.

Theorem 1. *Absent an identification strategy for the mediator, causal mediation estimates are at risk of selection bias. Suppose 1(1) holds, but 1(2) does not. Then CM estimands are contaminated by selection bias and group differences. Proof: see Appendix A.2.*

Below I present the relevant selection bias and group difference terms, omitting the conditional on $\boldsymbol{X}_i$ notation for brevity.
For the direct effect: CM estimand = ADE + selection bias + group differences.[7]

$$
\begin{aligned}
&\mathbb{E}_{D_i}\Big[\mathbb{E}\left[Y_i \mid Z_i = 1, D_i\right] - \mathbb{E}\left[Y_i \mid Z_i = 0, D_i\right]\Big] \\
&= \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] \\
&\quad + \mathbb{E}_{D_i = d'}\Big[\mathbb{E}\left[Y_i(0, D_i(Z_i)) \mid D_i(1) = d'\right] - \mathbb{E}\left[Y_i(0, D_i(Z_i)) \mid D_i(0) = d'\right]\Big] \\
&\quad + \mathbb{E}_{D_i = d'}\left[\Big(1 - \Pr\left(D_i(1) = d'\right)\Big)\begin{pmatrix}\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid D_i(1) = 1 - d'\right] \\ -\, \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid D_i(1) = d'\right]\end{pmatrix}\right]
\end{aligned}
$$

For the indirect effect: CM estimand = AIE + selection bias + group differences.

$$
\begin{aligned}
&\mathbb{E}_{Z_i}\left[\Big(\mathbb{E}\left[D_i \mid Z_i = 1\right] - \mathbb{E}\left[D_i \mid Z_i = 0\right]\Big) \times \Big(\mathbb{E}\left[Y_i \mid Z_i, D_i = 1\right] - \mathbb{E}\left[Y_i \mid Z_i, D_i = 0\right]\Big)\right] \\
&= \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] \\
&\quad + \Pr\left(D_i(1) = 1, D_i(0) = 0\right)\Big(\mathbb{E}\left[Y_i(Z_i, 0) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(Z_i, 0) \mid D_i = 0\right]\Big) \\
&\quad + \Pr\left(D_i(1) = 1, D_i(0) = 0\right) \times \\
&\quad \left[\begin{array}{l}\Big(1 - \Pr\left(D_i = 1\right)\Big)\begin{pmatrix}\mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \mid D_i = 1\right] \\ -\, \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \mid D_i = 0\right]\end{pmatrix} \\ -\left(\dfrac{1 - \Pr\left(D_i(1) = 1, D_i(0) = 0\right)}{\Pr\left(D_i(1) = 1, D_i(0) = 0\right)}\right)\begin{pmatrix}\mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \mid D_i(1) = 0 \text{ or } D_i(0) = 1\right] \\ -\, \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0)\right]\end{pmatrix}\end{array}\right]
\end{aligned}
$$

The selection bias terms come from systematic differences between the groups taking or refusing the mediator ($D_i = 1$ versus $D_i = 0$), differences not fully unexplained by $\boldsymbol{X}_i$. These

[7] The bias terms here mirror those in Heckman, Ichimura, Smith & Todd (1998), Angrist & Pischke (2009) for a single $D \to Y$ treatment effect, when D_i is not ignorable:

$$\mathbb{E}\left[Y_i \mid D_i = 1\right] - \mathbb{E}\left[Y_i \mid D_i = 0\right] = \text{ATE} + \underbrace{\Big(\mathbb{E}\left[Y_i(.,0) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(.,0) \mid D_i = 0\right]\Big)}_{\text{Selection Bias}} + \underbrace{\Pr\left(D_i = 0\right)\left(\text{ATT} - \text{ATU}\right)}_{\text{Group-differences Bias}}.$$

selection bias terms would equal zero if the mediator had been ignorable 1(2), but do not necessarily average to zero if not.

The group differences represent the fact that a matching approach gives an average effect on the treated group and, when selection-on-observables does not hold, this is systematically different from the average effect (Heckman et al. 1998). These terms are a non-parametric framing of the bias from controlling for intermediate outcomes, previously studied only in a linear setting (i.e., bad controls in Cinelli, Forney & Pearl 2024, or M-bias in Ding & Miratrix 2015).

The AIE group differences term is longer, because the indirect effect is comprised of the effect of D_i local to Z_i compliers.

$$\text{AIE} = \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] = \mathbb{E}\left[D_i(1) - D_i(0)\right] \underbrace{\mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i(1) = 1, D_i(0) = 0\right]}_{\text{Average } D \to Y \text{ effect among } D(Z) \text{ compliers}}$$

It is important to acknowledge the mediator compliers here, because the AIE is the $Z \to Y$ effect going through the $D(Z)$ channel, thus only refers to individuals pushed into mediator D_i by initial treatment Z_i. If we had been using a population average effect for $D \to Y$, then this is losing focus on the definition of the AIE; it is not about the causal effect $D \to Y$, it is about the causal effect $D(Z) \to Y$.

The group difference bias term in 1 arises because the selection-on-observables approach assumes that this complier average effect is equal to the population average effect, which does not hold true if the mediator is not ignorable.

2 Causal Mediation (CM) in Applied Settings

Unobserved confounding is particularly problematic when studying the mechanisms behind treatment effects. For example, in studying health gains from health insurance, we might expect that health gains came about because those with new insurance started visiting their healthcare provider more often, when in past they forewent using healthcare over financial concerns (Finkelstein et al. 2012). Applying conventional CM methods to investigate this expectation would be dismissing unobserved confounders for how often individuals visit healthcare providers, leading to biased results.

The wider population does not have one uniform bill of health; many people are born predisposed to ailments, due to genetic variation or other unrelated factors. These conditions can exist for years before being diagnosed. People with severe underlying conditions may visit healthcare providers more often than the rest of the population, to investigate or begin treating the ill–effects. It stands to reason that people with more serve underlying conditions may gain more from more often attending healthcare providers once given health insurance. These underlying causes for responding more to new access to health insurance cannot be controlled for by researchers, as researchers cannot hope to observe and control for health conditions that are yet to even be diagnosed. This means underlying health conditions are an unobserved confounder, and will bias estimates of the ADE and AIE in this setting.

In this section, I further develop the issue of selection on unobserved factors in a general CM setting. First, I show the non-parametric bias terms from Section 1 can be written as omitted variables bias in a regression framework. Second, I show how selection bias operates in a basic model for selection-into-mediator based on costs and benefits.

2.1 Regression Framework

Inference for CM effects can be written in a regression framework, showing how correlation between the error term and the mediator persistently biases estimates.

Start by writing potential outcomes $Y_i(.,.)$ as a sum of observed and unobserved factors, following the notation of Heckman & Vytlacil (2005). For each $z', d' = 0, 1$, put $\mu_{d'}(z'; \boldsymbol{X}_i) = \mathbb{E}\left[Y_i(z', d') \,|\, \boldsymbol{X}_i\right]$ and the corresponding error terms, $U_{d',i} = Y_i(z', d') - \mu_{d'}(z'; \boldsymbol{X}_i)$, so we have the following expressions:

$$Y_i(Z_i, 0) = \mu_0(Z_i; \boldsymbol{X}_i) + U_{0,i}, \quad Y_i(Z_i, 1) = \mu_1(Z_i; \boldsymbol{X}_i) + U_{1,i}.$$

In these terms, the ADE and AIE are represented as follows,

$$\begin{aligned} \text{ADE} &= \mathbb{E}\left[D_i\Big(\mu_1(1; \boldsymbol{X}_i) - \mu_1(0; \boldsymbol{X}_i)\Big) + (1 - D_i)\Big(\mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\Big)\right], \\ \text{AIE} &= \mathbb{E}\left[\Big(D_i(1) - D_i(0)\Big) \times \Big(\mu_1(Z_i; \boldsymbol{X}_i) - \mu_0(Z_i; \boldsymbol{X}_i) + U_{1,i} - U_{0,i}\Big)\right]. \end{aligned}$$

With this notation, observed data $Z_i, D_i, Y_i, \boldsymbol{X}_i$ have the following outcome equations — which characterise direct effects, indirect effects, and selection bias.

$$D_i = \phi + \overline{\pi} Z_i + \zeta(\boldsymbol{X}_i) + \eta_i \tag{3}$$

$$Y_i = \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) + \underbrace{(1 - D_i)\, U_{0,i} + D_i U_{1,i}}_{\text{Correlated error term.}} \tag{4}$$

First-stage (3) is identified, with $\phi + \zeta(\boldsymbol{X}_i)$ the intercept, and $\overline{\pi}$ the first-stage compliance rate (which may depend on $\boldsymbol{X}_i$). Second-stage (4) has the following definitions, and is not identified thanks to omitted variables bias. See Appendix A.3 for the derivation.

(a) $\alpha = \mathbb{E}\left[\mu_0(0; \boldsymbol{X}_i)\right]$ and $\varphi(\boldsymbol{X}_i) = \mu_0(0; \boldsymbol{X}_i) - \alpha$ are the intercept terms.

(b) $\beta = \mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)$ is the AIE local to $Z_i = 0$.

(c) $\gamma = \mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)$ is the ADE local to $D_i = 0$.

(d) $\delta = \mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - (\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i))$ is the average interaction effect.

(e) $(1 - D_i)\, U_{0,i} + D_i U_{1,i}$ is the disruptive error term.

The ADE and AIE are averages of these regression coefficients.

$$\text{ADE} = \mathbb{E}\left[\gamma + \delta D_i\right],$$
$$\text{AIE} = \mathbb{E}\left[\overline{\pi}\left(\beta + \delta Z_i + \widetilde{U}_i\right)\right], \qquad \text{with } \widetilde{U}_i = \underbrace{\mathbb{E}\left[U_{1,i} - U_{0,i} \,|\, \boldsymbol{X}_i, D_i(1) = 1, D_i(0) = 0\right]}_{\text{Unobserved complier gains.}}.$$

The ADE is a simple sum of the coefficients, while the AIE includes a group differences term because it only refers to $D(Z)$ compliers.

By construction, $\boldsymbol{U}_i := (U_{0,i}, U_{1,i})$ is an unobserved confounder. The regression estimates of β, γ, δ in second-stage (4) give unbiased estimates only if D_i is also conditionally ignorable: $D_i \perp\!\!\!\perp \boldsymbol{U}_i$. If not, then estimates of CM effects suffer from omitted variables bias from failing to adjust for the unobserved confounder, $\boldsymbol{U}_i$.

2.2 Selection on Costs and Benefits

CM is at risk of bias because $D_i \perp\!\!\!\perp (U_{0,i}, U_{1,i})$ is unlikely to hold in applied settings. A separate identification strategy could disrupt the selection-into-D_i based on unobserved factors, and lend credibility to the mediator ignorability assumption. Without it, bias will persist, given how we conventionally think of selection-into-treatment.

Consider a model where individual i selects into a mediator based on costs and benefits (in terms of outcome Y_i), after $Z_i, \boldsymbol{X}_i$ have been assigned. In a natural experiment setting, an external factor has disrupted individuals selecting Z_i by choice (thus Z_i is ignorable), but it has not disrupted the choice to take mediator (thus D_i is not ignorable). Write C_i for individual i's costs of taking mediator D_i, and $\mathbb{1}\{.\}$ for the indicator function. The Roy model has i taking the mediator if the benefits exceed the costs,

$$D_i\left(z'\right) = \mathbb{1}\left\{\underbrace{Y_i\left(z', 1\right) - Y_i\left(z', 0\right)}_{\text{Benefits}} \geq \underbrace{C_i}_{\text{Costs}}\right\}, \quad \text{for } z' = 0, 1. \tag{5}$$

The Roy model provides an intuitive framework for analysing selection mechanisms because it captures the fundamental economic principle of decision-making based on costs and benefits in terms of the outcome under study (Roy 1951, Heckman & Honore 1990). If the outcome Y_i is a measure of income, and the mediator a choice of taking education, then it models an individual choice to attend more education in terms of gaining a higher income compared to the costs.[8] This makes it particularly useful as a base case for CM, where selection-into-the mediator may be driven by private information (unobserved by the researcher). By using the Roy model as a benchmark, I explore the practical limits of the mediator ignorability assumption.

Decompose the costs into its mean and an error term, $C_i(Z_i) = \mu_C(Z_i; \boldsymbol{X}_i) + U_{C,i}$, to give a representation of Roy selection in terms of observed and unobserved factors,

$$D_i(z') = \mathbb{1}\left\{\mu_1(z'; \boldsymbol{X}_i) - \mu_0(z'; \boldsymbol{X}_i) - \mu_C(z'; \boldsymbol{X}_i) \geq U_{C,i} - \big(U_{1,i} - U_{0,i}\big)\right\}, \quad \text{for } z' = 0, 1.$$

If selection is Roy style, and the mediator is ignorable, then unobserved benefits play no part in selection. The only driver in differences in selection are differences in costs (and not benefits). If there are any unobserved benefits for selection-into-D_i unobserved to the researcher, then sequential ignorability cannot hold.

Proposition 1. *Suppose mediator selection follows a Roy model* (5)*, and selection is not fully explained by costs and observed gains. Then sequential ignorability does not hold.*

If there are any unobserved sources of gains, then sequential ignorability does not hold. This is an equivalence statement: selection based on costs and benefits is only consistent with mediator ignorability if the researcher observed every single source of mediator benefits. See Appendix A.4 for the proof.

This means than the vector of control variables $\boldsymbol{X}_i$ must be incredibly rich. Together, $\boldsymbol{X}_i$ and unobserved cost differences $U_{C,i}$ must explain selection-into-D_i one hundred percent. In the Roy model framework, however, individuals make decisions about mediator take-up based on gains, which the researcher may not observe fully. These unobservables are unlikely to be fully captured by an observed control set $\boldsymbol{X}_i$, except in very special cases (see e.g., the discussion in Angrist & Pischke 2009, Angrist 2022). In practice, the only way to believe in the mediator ignorability assumption is to study a setting where the researcher has two causal research designs, one for treatment Z_i and another for mediator D_i, at the same time. A simple addition of "we assume the mediator satisfies selection-on-observables" will not cut it here, and will lead to biased inference in practice.

[8] If the choice is made for a sum of outcomes, then a simple extension to a utility maximisation model maintains this same framework. See Heckman & Honore (1990), Eisenhauer, Heckman & Vytlacil (2015).

3 Solving Identification with a Control Function (CF)

If your goal is to estimate CM effects, and you could control for unobserved selection terms $U_{0,i}, U_{1,i}$, then you would. This ideal (but infeasible) scenario would yield unbiased estimates for the ADE and AIE. A Control Function (CF) approach takes this insight seriously, providing conditions to model the implied confounding by $U_{0,i}, U_{1,i}$, and then controlling for it.

The main problem is that second-stage regression equation (4) is not identified, because $U_{0,i}, U_{1,i}$ are unobserved and lead to omitted variables bias.

$$
\begin{aligned}
\mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i\right] \;=\; & \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\
& + (1 - D_i)\, \mathbb{E}\left[U_{0,i} \,|\, D_i = 0, \boldsymbol{X}_i\right] + D_i \mathbb{E}\left[U_{1,i} \,|\, D_i = 1, \boldsymbol{X}_i\right] \qquad (6)
\end{aligned}
$$

The CF approach models the contaminating terms in (6), avoiding the bias from omitting them in regression estimates. CF methods were first devised to correct for sample selection problems (Heckman 1974), and were extended to a general selection problem of the same form as Equation (6) (Heckman 1979). The approach works in the following manner: (1) assume that the variable of interest follows a selection model, where unexplained first-stage selection informs unobserved second-stage confounding; (2) extract information about unobserved confounding from the first-stage; and (3) incorporate this information as control terms in the second-stage equation to adjust for selection-into-mediator. Identification in CF methods typically relies on either distributional assumptions on the unobserved error terms, or an exclusion restriction for instrumental variables in the first-stage (or both). By explicitly accounting for the information contained in the first-stage selection model, CF methods enable consistent estimation of causal effects in the second-stage even when selection is driven by unobserved factors (Florens et al. 2008).

In the example of analysing health gains from health insurance (Finkelstein et al. 2012), a CF approach addresses the unobserved confounding from underlying health conditions. It does so by assuming that unobserved selection-into-frequent health care usage is informative for underlying health conditions, assuming people with more severe underlying conditions visit the doctor more often than those without. Then it uses this information in the second-stage estimation of how much the effect goes through increased healthcare usage, estimating the ADE and AIE.

3.1 Reidentifying Causal Mediation (CM) Effects

The following assumptions are sufficient to model the correlated error terms, identifying β, γ, δ in the second-stage regression (4), and thus both the ADE and AIE.

Assumption CF–1. Mediator monotonicity, conditional on $\boldsymbol{X}_i$.

$$\Pr\left(D_i(1) \geq D_i(0) \mid \boldsymbol{X}_i\right) = 1.$$

Assumption CF–1 is the monotonicity condition first used in an instrumental variables context (Imbens & Angrist 1994). Here, it is assuming that people respond to treatment, Z_i, by consistently taking or refusing the mediator D_i (always or never-takers), or taking the mediator D_i if and only if assigned to the treatment $Z_i = 1$ (mediator compliers). There are no mediator defiers.

The main implication of Assumption CF–1 is that selection-into-mediator can be written as a selection model with ordered threshold crossing values that describe selection-into-D_i (Vytlacil 2002).

$$D_i(z') = \mathbb{1}\left\{\psi\big(z'; \boldsymbol{X}_i\big) \geq V_i\right\}, \quad \text{for } z' = 0, 1$$

where V_i is a latent variable with continuous distribution and conditional cumulative density function $F_V(\,.\,|\boldsymbol{X}_i)$, and $\psi(\,.\,; \boldsymbol{X}_i)$ collects observed sources of mediator selection. V_i could be assumed to follow a known distribution; the canonical Heckman selection model assumes V_i is normally distributed (a "Heckit" model). The identification strategy here applies to the general case that the distribution of V_i is unknown, without parametric restrictions.

I focus on the equivalent transformed model of Heckman & Vytlacil (2005),

$$D_i(z') = \mathbb{1}\left\{\pi(z'; \boldsymbol{X}_i) \geq U_i\right\}, \quad \text{for } z' = 0, 1$$

where $U_i \coloneqq F_V\left(V_i \mid \boldsymbol{X}_i\right)$ follows a uniform distribution conditional, and $\pi(z'; \boldsymbol{X}_i) = F_V\big(\psi(z'; \boldsymbol{X}_i)\big) = \Pr\left(D_i = 1 \,|\, Z_i = z', \boldsymbol{X}_i\right)$ is the mediator propensity score. U_i are the unobserved mediator take-up costs. Note the maintained assumption that treatment Z_i is ignorable (conditional on $\boldsymbol{X}_i$) implies $Z_i \perp\!\!\!\perp U_i$ conditional on $\boldsymbol{X}_i$.

This selection model setup is equivalent to the monotonicity condition, and is importing a well-known equivalence result from the instrumental variables literature to the CM setting; this work is merely importing this result to the CM setting. The main conceptual difference is not assuming Z_i is a valid instrument for identifying complier $D \to Y$ effects; it is using the selection model representation to correct for selection bias. See Appendix A.5 for a validation of the general Vytlacil (2002) equivalence result in a CM setting, with conditioning covariates $\boldsymbol{X}_i$.

Assumption CF–2. Selection on mediator costs and benefits.

$$\text{Cov}\left(U_i,\, U_{0,i}\right),\ \text{Cov}\left(U_i,\, U_{1,i}\right) \neq 0.$$

Assumption CF–2 is stating that unobserved selection in mediator take-up (U_i) informs second-stage confounding, when refusing or taking the mediator ($U_{0,i}$ and $U_{1,i}$).

This is a strong assumption, and will not hold in all examples. If people had been deciding to take D_i by a Roy model, then this assumption holds because $V_i = U_{C,i} - \big(U_{1,i} - U_{0,i}\big)$. Individuals could be making decisions based on other outcomes, but as long as mediator costs and benefits guide at least part of this decision (i.e., bounded away from zero), then this assumption will hold.

For notation purposes, suppose the vector of control variables $\boldsymbol{X}_i$ has at least two entries; denote $\boldsymbol{X}_i^{\text{IV}}$ as one entry in the vector, and $\boldsymbol{X}_i^{-}$ as the remaining.

Assumption CF–3. Mediator take-up cost instrument.

$$\boldsymbol{X}_i^{\text{IV}} \text{ satisfies } \frac{\partial}{\partial \boldsymbol{X}_i^{\text{IV}}}\Big\{\mu_1(z', \boldsymbol{X}_i) - \mu_0(z', \boldsymbol{X}_i)\Big\} = 0 < \frac{\partial}{\partial \boldsymbol{X}_i^{\text{IV}}}\Big\{\mathbb{E}\left[D_i(z') \,|\, \boldsymbol{X}_i\right]\Big\}, \text{ for } z' = 0, 1.$$

Assumption CF–3 is requiring at least one control variable guides selection-into-D_i — an Instrumental Variable (IV). It assumes an instrument exists, which satisfies an exclusion restriction (i.e., not impacting mediator gains $\mu_1 - \mu_0$), and has a non-zero influence on the mediator (i.e., strong IV first-stage). The exclusion restriction is untestable, and must be guided by domain-specific knowledge; IV first-stage strength is testable, and must be justified with data by methods common in the IV literature.

This assumption identifies the mediator propensity score separately from the direct and indirect effects, avoiding indeterminacy in the second-stage outcome equation. While not technically required for identification, it avoids relying entirely on an assumed distribution for unobserved error terms (and bias from inevitably breaking this assumption). The most compelling example of a mediator IV is using data on the cost of mediator take-up as a first-stage IV, if it varies between individuals for unrelated reasons and is strong in explaining mediator take-up.

Proposition 2. *If assumptions CF–1, CF–2, CF–3 hold, then second-stage regression equation (4) is identified with a CF adjustment.*

$$\begin{aligned}\mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i\right] \;&=\; \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\ &\qquad + \rho_0 \left(1 - D_i\right) \lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big) + \rho_1 D_i \lambda_1\big(\pi(Z_i; \boldsymbol{X}_i)\big),\end{aligned}$$

where λ_0, λ_1 are the Control Functions (CFs), ρ_0, ρ_1 are linear parameters, and mediator *propensity score* $\pi(z'; \boldsymbol{X}_i)$ *is identified from the first-stage* (3)*. Proof: see Appendix A.6.*

The CFs are functions which measure unobserved mediator gains, for those with unobserved mediator costs above or below a propensity score value. Following the IV notation of Kline & Walters (2019), put $\mu_V = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)\right]$, to give the following representation for the CFs:

$$\lambda_0\big(p'\big) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, p' < U_i\right],$$
$$\lambda_1\big(p'\big) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, U_i \leq p'\right] = -\lambda_0\big(p'\big)\left(\frac{1-p'}{p'}\right), \text{ for } p' \in (0,1).$$

If we are using the canonical Heckman selection model, we assume the error term follows a normal distribution, so that λ_0, λ_1 are the inverse Mills ratio. Alternatively, λ_0, λ_1 could have other definitions following the assumed distribution of the error terms (see e.g, Wooldridge 2015). If we do not know what distribution class the errors follow, then λ_0, λ_1 can be estimated separately with semi-parametric methods to avoid relying on parametric assumptions.

Theorem CF. If assumptions CF–1, CF–2, CF–3 hold, the ADE and AIE are identified as a function of the parameters in Proposition 2.

$$\text{ADE} = \mathbb{E}\left[\gamma + \delta D_i\right],$$
$$\text{AIE} = \mathbb{E}\left[\overline{\pi}\left(\beta + \delta Z_i + \underbrace{(\rho_1 - \rho_0)\,\Gamma\big(\pi(0; \boldsymbol{X}_i),\, \pi(1; \boldsymbol{X}_i)\big)}_{\text{Complier adjustment}}\right)\right]$$

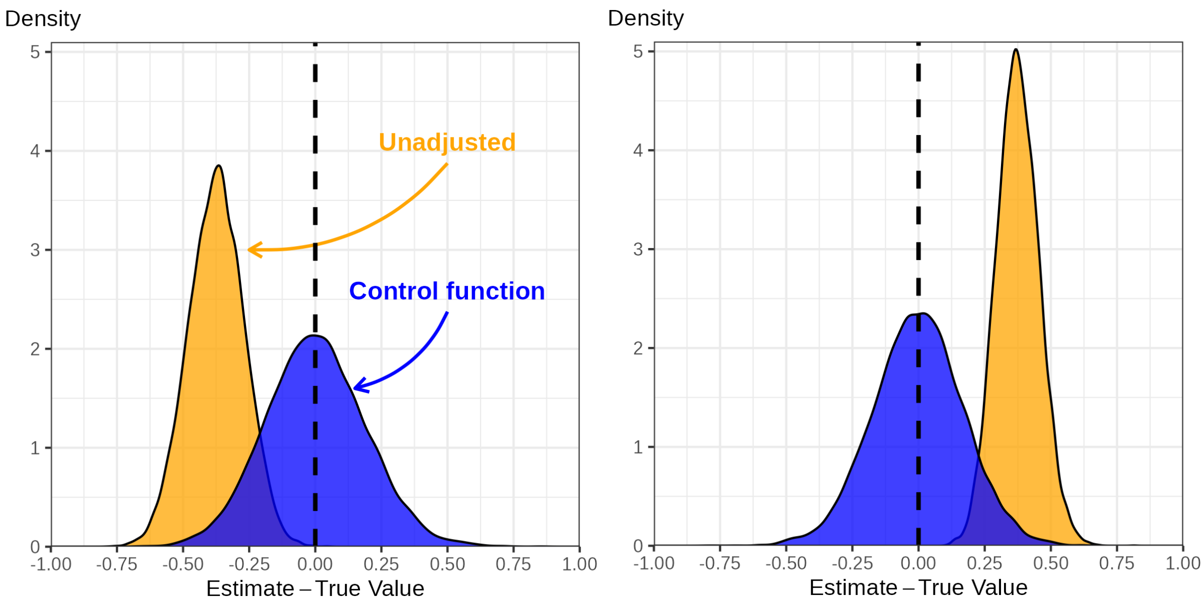
where $\Gamma\left(p, p'\right) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, p < U_i \leq p'\right] = \frac{p'\lambda_1(p') - p\lambda_1(p)}{p'-p}$ is the average unobserved net gains for those with unobserved costs between $p < p'$, and $\overline{\pi} = \pi(1; \boldsymbol{X}_i) - \pi(0; \boldsymbol{X}_i)$ is the mediator complier score. Proof: see Appendix A.7.

This theorem provides a solution to the identification problem for CM effects when facing selection; rather than assuming away selection problems, it explicitly models them. The ADE is straightforward to calculate as an average of the direct effect parameters, while the AIE also includes an adjustment for unobserved complier gains to the mediator. Again, this is because the AIE only refers to individuals who were induced by treatment Z_i into taking mediator D_i (mediator compliers). The CFs allow us to measure both selection bias and complier differences, purging persistent bias in CM effect analyses.

Figure 2 shows how a CF adjustment corrects unadjusted CM effect estimates. In a simulation with selection-into-mediator based on unobserved error terms, the CF adjustment pushes conventional CM estimates back to the true value.

Figure 2: Simulated Distribution of CM Effect Estimates, Relative to True Value.

(a) $\widehat{\text{ADE}} - \text{ADE}$. **(b)** $\widehat{\text{AIE}} - \text{AIE}$.



Note: These figures show the empirical density of point estimates minus the true average effect, for 10,000 different datasets generated from a Roy model with correlated normally distributed error terms (further described in Section 4). The black dashed line is the true value; orange is the distribution of conventional CM estimates from two-stage OLS (Imai et al. 2010), and blue estimates with a two-stage Heckman selection adjustment.

4 Control Function (CF) Estimation of CM Effects

This paper only considers a modern two-stage approach to selection models — also known as a control function approach. Indeed, the conventional approach to estimating CM effects already uses a two-stage estimation procedure to estimate the first and second-stages. It is a simple, and intuition addition to include a two-stage control function adjustment to CM estimation.

Go through the steps of a Heckman selection model, and the corresponding SEs + reference $\sqrt{n}$-consistency. Note Nancy Heckamn (1986) gives $\sqrt{n}$-consistency of splines. The two-stage semi-parametric procedure of Robinson (1988) is not appropriate, as we want the $\lambda_1(.)$ function, too.

4.1 Simulation Evidence

The following simulation gives an example to show how this method works in practice. Suppose data observed to the researcher $Z_i, D_i, Y_i, \boldsymbol{X}_i$ are drawn from the following data generating processes, for $i = 1, \dots, N$.

$$Z_i \sim \text{Binom}\,(0.5)\,, \quad \boldsymbol{X}_i^- \sim N(4,1), \quad \boldsymbol{X}_i^{\text{IV}} \sim \text{Binom}\,(0.5)\,,$$
$$(U_{0,i}, U_{1,i}) \sim \text{BivariateNormal}\,(0,0,\sigma_0,\sigma_1,\rho)\,, \quad U_{C,i} \sim N(0,0.5).$$

$N = 10{,}000$ allows the large sample properties of the approach to operate; indeed, smaller sample sizes may not.

Suppose each i chooses to take mediator D_i by a Roy model, with following mean definitions for each $z', d' = 0, 1$.

$$D_i(z') = \mathbb{1}\left\{Y_i(z',1) - Y_i(z',0) \geq C_i\right\},$$
$$\mu_{d'}\left(z'; \boldsymbol{X}_i\right) = \boldsymbol{X}_i^- + \left(z' + d' + z'd'\right), \quad \mu_C\left(z'; \boldsymbol{X}_i\right) = 3z' + \boldsymbol{X}_i^- - \boldsymbol{X}_i^{\text{IV}}.$$

Following Section 2, these data have the following first and second-stage equations:

$$D_i = \mathbb{1}\left\{-3Z_i - \boldsymbol{X}_i^{\text{IV}} + \boldsymbol{X}_i^- \geq U_{C,i} - \Big(U_{1,i} - U_{0,i}\Big)\right\},$$
$$Y_i = Z_i + D_i + Z_i D_i + \boldsymbol{X}_i^- + (1 - D_i)\,U_{0,i} + D_i U_{1,i}.$$

Z_i has an effect on outcome Y_i, and it operates partially through mediator D_i. Outcome mean $\mu_{D_i}(Z_i; .)$ contains an interaction term, $Z_i D_i$, so while both Z_i and D_i have constant partial effects, the ATE depends on how many i choose to take the mediator. In this simulation $\Pr\left(D_i = 1\right) = 0.437$, and 65.29% of the sample are mediator compliers (where $D_i(1) = 1$ and $D_i(0) = 0$). This gives an ATE ($Z \to Y$) value of 2.58, ADE 1.44, and AIE 1.13, respectively.[9]

After Z_i is assigned, i chooses to take mediator D_i by considering the costs and benefits — which vary based on Z_i, demographic controls $\boldsymbol{X}_i$, and the (non-degenerate) unobserved error terms $U_{i,0}, U_{1,i}$. As a result, sequential ignorability does not hold; the mediator is not conditionally ignorable. Thus, a standard OLS (selection-on-observables) approach to CM does not give an estimate for how much of the $Z \to Y$ ATE goes through mediator D. Instead, the OLS approach gives biased inference. The bias in OLS estimates comes from the unobserved error terms being related.

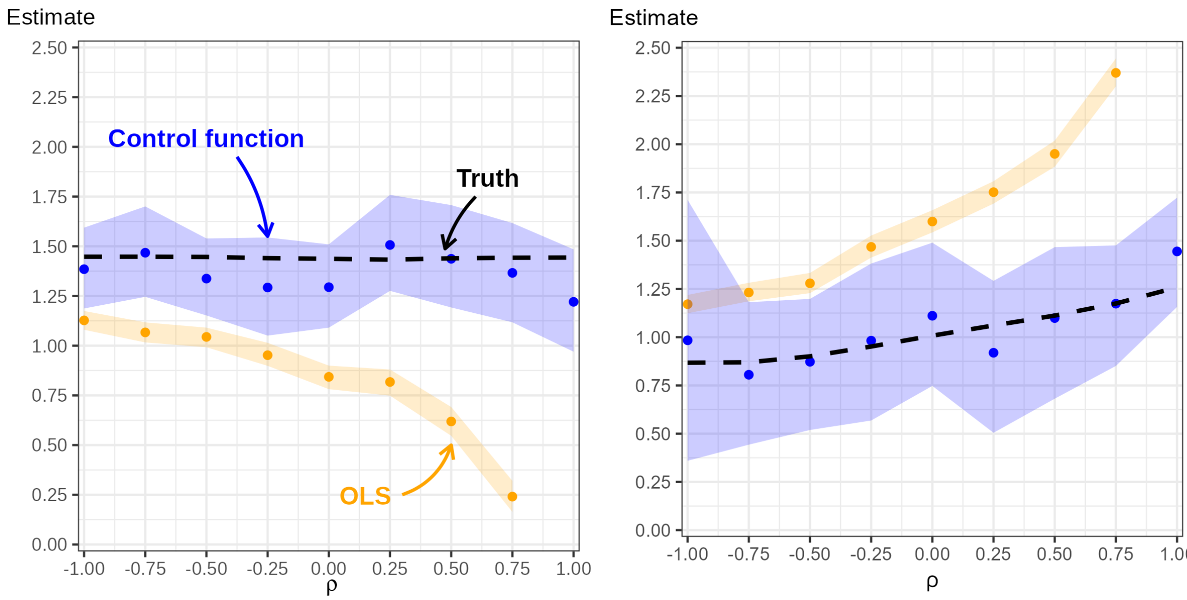
The CF estimator here uses the inverse Mills ratio as the CFs (i.e., a Heckman selection model); in-progress work develops a constrained semi-parametric approach to model λ_0 (and thus λ_1) to avoid distributional assumptions. Figure 2 shows the distribution of simulated point estimates in this simulation, showing OLS against the CF approach. The OLS approach

[9]Note that ATE = ADE + AIE in this setting. $\Pr\left(Z_i = 1\right) = 0.5$ ensures this equality, but it is not guaranteed in general.

implicitly assumes that the mediator is ignorable (when it is not), so its point estimates under and over-estimate the true ADE and AIE, respectively. The distance between the OLS estimates and the true values are the underlying bias terms derived in Theorem 1. In this data generating process, the OLS confidence interval do not overlap the true values for any standard level of significance. The CF approach exhibits bias, though the 95% confidence intervals cover the truth.

Figure 3: Point Estimates of CM Effects, OLS versus CF, varying ρ values with $\sigma_0 = 1, \sigma_1 = 2$ fixed.

(a) ADE. **(b)** AIE.



Note: These figures show the OLS and CF point estimates of the ADE and AIE, for $N = 10,000$ sample size. The black dashed line is the true value, points are points estimates from data simulated with a given ρ value and $\sigma_0 = 1, \sigma_1 = 2$, and shaded regions are the 95% confidence intervals from 1,000 bootstraps each. Orange represents OLS estimates, blue the CF approach. The true AIE values vary with ρ, because $D_i(Z_i)$ compliers have higher average values of $U_{1,i} - U_{0,i}$ with greater ρ values.

The error terms determine the bias in OLS estimates of the ADE and AIE, so the bias varies for different values of the error-term parameters $\rho \in [-1, 1]$ and $\sigma_0, \sigma_1 \geq 0$.[10] Figure 3 shows CF estimates against estimates calculated by standard OLS, showing 95% confidence intervals calculated from 1,000 bootstraps. The point estimates of the CF do not exactly equal the true values, as they are estimates from one simulation (not averages across many simulations, as in Figure 2). The CF approach improves on OLS estimates by correcting for

[10]Indeed, this setting has error terms following a bivariate normal distribution, so the canonical Heckman (1974) selection model would produce the most efficient estimates by maximum likelihood. The CF approach avoids this assumption, and bias from breaking it, by relying on an instrument.

bias, with confidence regions overlapping the true values.[11,12] This correction did not come for free: the standard errors are significantly greater in a CF approach than OLS. Standard errors on the AIE are larger than those for the ADE, because the AIE estimates are first-stage times second-stage estimates, so standard errors account for uncertainty in both estimates multiplied. In this manner, this simulation shows the pros and cons of using the CF approach to estimating CM effects in practice.

5 Summary and Concluding Remarks

This paper has studied a selection-on-observables approach to CM in a natural experiment setting. I have shown the pitfalls of using the most popular methods for estimating direct and indirect effects without a clear case for the mediator being ignorable. Using the Roy model as a benchmark, a mediator is unlikely to be ignorable in natural experiment settings, and the bias terms likely crowd out inference regarding CM effects.

This paper has contributed to the growing CM literature in economics, integrating labour economic theory for selection-into-treatment as a way of judging the credibility of conventional CM analyses. It has drawn on the classic literature, and pointed to already-in-use selection models/control function methods as a compelling way of estimating direct and indirect effects in a natural experiment setting. Further research could build on this approach by suggesting efficiency improvements, adjustments for common statistical irregularities (say, cluster dependence), or integrating the selection model/control function as an additional robustness in the growing double robustness literature (Farbmacher, Huber, Laffërs, Langen & Spindler 2022, Bia, Huber & Laffërs 2024).

This paper does not provide a blanket endorsement for applied researchers to use CM methods. The structural assumptions are strong, and design-based inference requires an instrument for mediator take-up; if the assumptions are broken, then selection-adjusted estimates of CM effects will also be biased, and will not improve on the selection-on-observables approach. And yet, there are likely settings in which the structural assumptions are credible. Mediator monotonicity aligns well with economic theory in many cases, and it is plausible for researchers to study big data settings with external variation in mediator take-up costs. In these cases, this paper opens the door to identifying mechanisms behind treatment effects

[11]The code behind this simulation estimates the first-stage with an interacted OLS specification. The second-stage is an OLS specification, including the CFs. The in-progress semi-parametric approach uses a spline specification for λ_1 in the $D_i = 1$ sample, and λ_0 in the $D_i = 0$ sample, composing a moment restricted estimate of λ_1 from these.

[12]In the appendix, Figure A1 shows the same simulation while varying σ_1, with fixed $\sigma_0 = 1, \rho = 0.5$. The conclusion is the same as for varying the correlation coefficient, ρ, in Figure 3.

in natural experiment settings.

References

Angrist, J. D. (1998), 'Estimating the labor market impact of voluntary military service using social security data on military applicants', *Econometrica* **66**(2), 249–288. 6

Angrist, J. D. (2022), 'Empirical strategies in economics: Illuminating the path from cause to effect', *Econometrica* **90**(6), 2509–2539. 11

Angrist, J. D. & Pischke, J.-S. (2009), *Mostly harmless econometrics: An empiricist's companion*, Princeton university press. 7, 11

Bia, M., Huber, M. & Lafférs, L. (2024), 'Double machine learning for sample selection models', *Journal of Business & Economic Statistics* **42**(3), 958–969. 19

Blackwell, M., Ma, R. & Opacic, A. (2024), 'Assumption smuggling in intermediate outcome tests of causal mechanisms', *arXiv preprint arXiv:2407.07072* . 2

Cinelli, C., Forney, A. & Pearl, J. (2024), 'A crash course in good and bad controls', *Sociological Methods & Research* **53**(3), 1071–1104. 8

Deuchert, E., Huber, M. & Schelker, M. (2019), 'Direct and indirect effects based on difference-in-differences with an application to political preferences following the vietnam draft lottery', *Journal of Business & Economic Statistics* **37**(4), 710–720. 2

Ding, P. & Miratrix, L. W. (2015), 'To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias', *Journal of Causal Inference* **3**(1), 41–57. 8

Eisenhauer, P., Heckman, J. J. & Vytlacil, E. (2015), 'The generalized roy model and the cost-benefit analysis of social programs', *Journal of Political Economy* **123**(2), 413–443. 11

Farbmacher, H., Huber, M., Lafférs, L., Langen, H. & Spindler, M. (2022), 'Causal mediation analysis with double machine learning', *The Econometrics Journal* **25**(2), 277–300. 19

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. & Group, O. H. S. (2012), 'The oregon health insurance experiment: Evidence from the first year*', *The Quarterly Journal of Economics* **127**(3), 1057–1106.
URL: *https://doi.org/10.1093/qje/qjs020* 1, 5, 8, 12

Florens, J.-P., Heckman, J. J., Meghir, C. & Vytlacil, E. (2008), 'Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects', *Econometrica* **76**(5), 1191–1206. 3, 12

Flores, C. A. & Flores-Lagunes, A. (2009), 'Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness'. 2

Frölich, M. & Huber, M. (2017), 'Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **79**(5), 1645–1666. 2, 3

Heckman, J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica: journal of the econometric society* pp. 679–694. 12, 18

Heckman, J., Ichimura, H., Smith, J. & Todd, P. (1998), 'Characterizing selection bias using experimental data', *Econometrica* **66**(5), 1017–1098. 7, 8

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica: Journal of the econometric society* pp. 153–161. 3, 12, 34

Heckman, J. J. & Honore, B. E. (1990), 'The empirical content of the roy model', *Econometrica: Journal of the Econometric Society* pp. 1121–1149. 3, 10, 11

Heckman, J. J. & Pinto, R. (2015), 'Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs', *Econometric reviews* **34**(1-2), 6–31. 2, 3

Heckman, J. J. & Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation 1', *Econometrica* **73**(3), 669–738. 3, 9, 13, 31

Heckman, J. & Navarro-Lozano, S. (2004), 'Using matching, instrumental variables, and control functions to estimate economic choice models', *Review of Economics and statistics* **86**(1), 30–57. 3

Heckman, J., Pinto, R. & Savelyev, P. (2013), 'Understanding the mechanisms through which an influential early childhood program boosted adult outcomes', *American economic review* **103**(6), 2052–2086. 3

Huber, M. (2020), 'Mediation analysis', *Handbook of labor, human resources and population economics* pp. 1–38. 2

Huber, M., Hsu, Y.-C., Lee, Y.-Y. & Lettry, L. (2020), 'Direct and indirect effects of continuous treatments based on generalized propensity score weighting', *Journal of Applied Econometrics* **35**(7), 814–840. 4

Imai, K., Keele, L. & Yamamoto, T. (2010), 'Identification, inference and sensitivity analysis for causal mediation effects', *Statistical Science* pp. 51–71. 2, 3, 5, 6, 16, 23, 28, 35

Imai, K., Tingley, D. & Yamamoto, T. (2013), 'Experimental designs for identifying causal mechanisms', *Journal of the Royal Statistical Society Series A: Statistics in Society* **176**(1), 5–51. 3

Imbens, G. & Angrist, J. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475. 4, 13

Kline, P. & Walters, C. R. (2019), 'On heckits, late, and numerical equivalence', *Econometrica* **87**(2), 677–696. 3, 15, 31, 33, 34

Kwon, S. & Roth, J. (2024), 'Testing mechanisms', *arXiv preprint arXiv:2404.11739* . 2

Ludwig, J., Kling, J. R. & Mullainathan, S. (2011), 'Mechanism experiments and policy evaluations', *Journal of economic Perspectives* **25**(3), 17–38. 3

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. 35

Robins, J. (1986), 'A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect', *Mathematical modelling* **7**(9-12), 1393–1512. 6

Roy, A. D. (1951), 'Some thoughts on the distribution of earnings', *Oxford economic papers* **3**(2), 135–146. 3, 10

Słoczyński, T. (2022), 'Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights', *Review of Economics and Statistics* **104**(3), 501–509. 6

Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. (2014), 'Mediation: R package for causal mediation analysis', *Journal of statistical software* **59**, 1–38. https://doi.org/10.18637/jss.v059.i05. 35

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**(1), 331–341. 3, 13, 30

Wang, W. & Yan, J. (2021), 'Shape-restricted regression splines with r package splines2.', *Journal of Data Science* **19**(3). 35

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019), 'Welcome to the tidyverse', *Journal of Open Source Software* **4**(43), 1686. https://doi.org/10.21105/joss.01686. 35

Wooldridge, J. M. (2015), 'Control function methods in applied econometrics', *Journal of Human Resources* **50**(2), 420–445. 3, 15

A Supplementary Appendix

This section is for supplementary information, and validation of presented propositions and theorems. It is not meant for publication.

Any comments or suggestions may be sent to me at seh325@cornell.edu, or raised as an issue on the Github project, https://github.com/shoganhennessy/mediation-natural-experiment.

A.1 Identification in Causal Mediation

Imai et al. (2010, Theorem 1) states that the ADE and AIE are identified under sequential ignorability, at each level of $Z_i = 0, 1$. For $z' = 0, 1$:

$$\mathbb{E}\left[Y_i(1, D_i(z')) - Y_i(0, D_i(z'))\right] = \int\int \Big(\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i, \boldsymbol{X}_i\right] \Big) dF_{D_i \,|\, Z_i = z', \boldsymbol{X}_i} dF_{\boldsymbol{X}_i},$$

$$\mathbb{E}\left[Y_i(z', D_i(1)) - Y_i(z', D_i(0))\right] = \int\int \mathbb{E}\left[Y_i \,|\, Z_i = z', D_i, \boldsymbol{X}_i\right] \Big(dF_{D_i \,|\, Z_i = 1, \boldsymbol{X}_i} - dF_{D_i \,|\, Z_i = 0, \boldsymbol{X}_i} \Big) dF_{\boldsymbol{X}_i}.$$

I focus on the averages, which are identified by consequence of the above.

$$\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] = \mathbb{E}_{Z_i}\left[\mathbb{E}\left[Y_i(1, D_i(z')) - Y_i(0, D_i(z')) \,|\, Z_i = z'\right]\right]$$
$$\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] = \mathbb{E}_{Z_i}\left[\mathbb{E}\left[Y_i(z', D_i(1)) - Y_i(z', D_i(0)) \,|\, Z_i = z'\right]\right]$$

My estimand for the ADE is a simple rearrangement of the above. The estimand for the AIE relies on a different sequence, relying on (1) sequential ignorability, (2) conditional monotonicity. These give (1) identification equivalence of AIE local to cpmpliers conditional on $\boldsymbol{X}_i$ and AIE conditional on $\boldsymbol{X}_i$, LAIE = AIE, (2) identification of the complier score.

$$\begin{aligned}
&\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \,|\, \boldsymbol{X}_i\right] \\
&= \Pr\left(D_i(1) = 1, D_i(0) = 0 \,|\, \boldsymbol{X}_i\right) \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right] \\
&= \Pr\left(D_i(1) = 1, D_i(0) = 0 \,|\, \boldsymbol{X}_i\right) \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, \boldsymbol{X}_i\right] \\
&= \Pr\left(D_i(1) = 1, D_i(0) = 0 \,|\, \boldsymbol{X}_i\right) \Big(\mathbb{E}\left[Y_i \,|\, Z_i, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i, D_i = 0, \boldsymbol{X}_i\right] \Big) \\
&= \Big(\mathbb{E}\left[D_i \,|\, Z_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[D_i \,|\, Z_i = 0, \boldsymbol{X}_i\right] \Big) \Big(\mathbb{E}\left[Y_i \,|\, Z_i, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i, D_i = 0, \boldsymbol{X}_i\right] \Big)
\end{aligned}$$

Monotonicity is not technically required for the above. Breaking monotonicity would not change the identification in any of the above; it would be the same except replacing the complier score with a complier/defier score, $\Pr\left(D_i(1) \neq D_i(0) \,|\, \boldsymbol{X}_i\right) = \mathbb{E}\left[D_i \,|\, Z_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[D_i \,|\, Z_i = 0, \boldsymbol{X}_i\right]$.

A.2 Bias in Mediation Estimates

Suppose that Z_i is ignorable conditional on $\boldsymbol{X}_i$, but D_i is not.

A.2.1 Bias in Direct Effect Estimates

To show that the conventional approach to mediation gives an estimate for the ADE with selection and group difference-bias, start with the components of the conventional estimands. This proof starts with the relevant expectations, conditional on a specific value of $\boldsymbol{X}_i$. For each $d' = 0, 1$.

$$\begin{aligned}
\mathbb{E}\left[Y_i \mid Z_i = 1, D_i = d', \boldsymbol{X}_i\right] =& \mathbb{E}\left[Y_i(1, D_i(Z_i)) \mid D_i(1) = d', \boldsymbol{X}_i\right], \\
\mathbb{E}\left[Y_i \mid Z_i = 0, D_i = d', \boldsymbol{X}_i\right] =& \mathbb{E}\left[Y_i(0, D_i(Z_i)) \mid D_i(0) = d', \boldsymbol{X}_i\right]
\end{aligned}$$

And so,

$$\begin{aligned}
&\mathbb{E}\left[Y_i \mid Z_i = 1, D_i = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \mid Z_i = 0, D_i = d', \boldsymbol{X}_i\right] \\
=& \mathbb{E}\left[Y_i(1, D_i(Z_i)) \mid D_i(1) = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(0, D_i(Z_i)) \mid D_i(0) = d', \boldsymbol{X}_i\right] \\
=& \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid D_i(1) = d', \boldsymbol{X}_i\right] \\
&+ \mathbb{E}\left[Y_i(0, D_i(Z_i)) \mid D_i(1) = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(0, D_i(Z_i)) \mid D_i(0) = d', \boldsymbol{X}_i\right].
\end{aligned}$$

The final term is a sum of the ADE, conditional on $D_i(1) = d'$, and a selection bias term — difference in baseline outcomes between the (partially overlapping) groups for whom $D_i(1) = d'$ and $D_i(0) = d'$.

To reach the final term, note the following.

$$\begin{aligned}
&\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid \boldsymbol{X}_i\right] \\
=& \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid D_i(1) = d', \boldsymbol{X}_i\right] \\
&+ \Big(1 - \Pr\left(D_i(1) = d' \mid \boldsymbol{X}_i\right)\Big) \begin{pmatrix} \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid D_i(1) = d', \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid D_i(1) = 1 - d', \boldsymbol{X}_i\right] \end{pmatrix}
\end{aligned}$$

The second term is the difference between the ADE and LADE local to relevant complier groups.

Collect everything together, as follows.

$$
\begin{aligned}
&\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = d', \boldsymbol{X}_i\right] \\
=&\underbrace{\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, \boldsymbol{X}_i\right]}_{\text{ADE, conditional on } \boldsymbol{X}_i} \\
&+ \underbrace{\mathbb{E}\left[Y_i(0, D_i(Z_i)) \,|\, D_i(1) = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(0, D_i(Z_i)) \,|\, D_i(0) = d', \boldsymbol{X}_i\right]}_{\text{Selection bias}} \\
&+ \underbrace{\Big(1 - \Pr\left(D_i(1) = d' \,|\, \boldsymbol{X}_i\right)\Big) \begin{pmatrix} \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, D_i(1) = 1 - d', \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, D_i(1) = d', \boldsymbol{X}_i\right] \end{pmatrix}}_{\text{group difference-bias}}
\end{aligned}
$$

The proof is achieved by applying the expectation across $D_i = d'$, and $\boldsymbol{X}_i$.

A.2.2 Bias in Indirect Effect Estimates

To show that the conventional approach to mediation gives an estimate for the AIE with selection and group difference-bias, start with the definition of the ADE — the direct effect among compliers times the size of the complier group.

This proof starts with the relevant expectations, conditional on a specific value of $\boldsymbol{X}_i$.

$$
\begin{aligned}
&\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \,|\, \boldsymbol{X}_i\right] \\
=& \Pr\left(D_i(1) = 1, D_i(0) = 0 \,|\, \boldsymbol{X}_i\right) \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right]
\end{aligned}
$$

When D_i is not ignorable, the bias comes from estimating the second term, $\mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right]$.

For each $z' = 0, 1$.

$$
\begin{aligned}
\mathbb{E}\left[Y_i \,|\, Z_i = z', D_i = 1, \boldsymbol{X}_i\right] =& \mathbb{E}\left[Y_i(z', 1) \,|\, D_i = 1, \boldsymbol{X}_i\right], \\
\mathbb{E}\left[Y_i \,|\, Z_i = z', D_i = 0, \boldsymbol{X}_i\right] =& \mathbb{E}\left[Y_i(z', 0) \,|\, D_i = 0, \boldsymbol{X}_i\right]
\end{aligned}
$$

So compose the CM estimand, as follows.

$$
\begin{aligned}
&\mathbb{E}\left[Y_i \,|\, Z_i = z', D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = z', D_i = 0, \boldsymbol{X}_i\right] \\
=&\mathbb{E}\left[Y_i(z', 1) \,|\, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(z', 0) \,|\, D_i = 0, \boldsymbol{X}_i\right] \\
=&\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i = 1, \boldsymbol{X}_i\right] + \mathbb{E}\left[Y_i(z', 0) \,|\, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(z', 0) \,|\, D_i = 0, \boldsymbol{X}_i\right]
\end{aligned}
$$

The final term is a sum of the AIE, among the treated group $D_i = 1$, and a selection bias term — difference in baseline terms between the groups $D_i = 1$ and $D_i = 0$.

The AIE is the direct effect among compliers times the size of the complier group, so we need to compensate for the difference between the treated group $D_i = 1$ and complier group $D_i(1) = 1, D_i(0) = 0$.

Start with the difference between treated group's average and overall average.

$$\begin{aligned}
&\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i = 1, \boldsymbol{X}_i\right] \\
=&\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid \boldsymbol{X}_i\right] \\
&+ \Big(1 - \Pr\left(D_i = 1 \mid \boldsymbol{X}_i\right)\Big) \begin{pmatrix} \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i = 1, \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i = 0, \boldsymbol{X}_i\right] \end{pmatrix}
\end{aligned}$$

Then the difference between the compliers' average and the overall average.

$$\begin{aligned}
&\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right] \\
=&\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid \boldsymbol{X}_i\right] \\
&+ \frac{1 - \Pr\left(D_i(1) = 1, D_i(0) = 0 \mid \boldsymbol{X}_i\right)}{\Pr\left(D_i(1) = 1, D_i(0) = 0 \mid \boldsymbol{X}_i\right)} \begin{pmatrix} \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i(1) = 0 \text{ or } D_i(0) = 1, \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid \boldsymbol{X}_i\right] \end{pmatrix}
\end{aligned}$$

Collect everything together, as follows.

$$\begin{aligned}
&\mathbb{E}\left[Y_i \mid Z_i = z', D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \mid Z_i = z', D_i = 0, \boldsymbol{X}_i\right] \\
=&\underbrace{\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right]}_{\text{AIE among compliers, conditional on } \boldsymbol{X}_i, Z_i = z'} \\
&+ \underbrace{\mathbb{E}\left[Y_i(z', 0) \mid D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(z', 0) \mid D_i = 0, \boldsymbol{X}_i\right]}_{\text{Selection bias}} \\
&+ \underbrace{\begin{bmatrix} \Big(1 - \Pr\left(D_i = 1 \mid \boldsymbol{X}_i\right)\Big) \begin{pmatrix} \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i = 1, \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i = 0, \boldsymbol{X}_i\right] \end{pmatrix} \\ - \frac{1 - \Pr\left(D_i(1) = 1, D_i(0) = 0 \mid \boldsymbol{X}_i\right)}{\Pr\left(D_i(1) = 1, D_i(0) = 0 \mid \boldsymbol{X}_i\right)} \begin{pmatrix} \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i(1) = 0 \text{ or } D_i(0) = 1, \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid \boldsymbol{X}_i\right] \end{pmatrix} \end{bmatrix}}_{\text{group difference-bias}}
\end{aligned}$$

The proof is finally achieved by multiplying by the complier score, $\Pr\left(D_i(1) = 1, D_i(0) = 0 \mid \boldsymbol{X}_i\right)$ $= \mathbb{E}\left[D_i \mid Z_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[D_i \mid Z_i = 0, \boldsymbol{X}_i\right]$, then applying the expectation across $Z_i = z'$, and $\boldsymbol{X}_i$.

A.3 A Regression Framework for Direct and Indirect Effects

Put $\mu_{d'}(z'; \boldsymbol{X}) = \mathbb{E}\left[Y_i(z', d') \,|\, \boldsymbol{X}\right]$ and $U_{d',i} = Y_i(z', d') - \mu_{d'}(z'; \boldsymbol{X})$ for each $z', d' = 0, 1$, so we have the following expressions:

$$Y_i(Z_i, 0) = \mu_0(Z_i; \boldsymbol{X}_i) + U_{0,i}, \quad Y_i(Z_i, 1) = \mu_1(Z_i; \boldsymbol{X}_i) + U_{1,i}.$$

$U_{0,i}, U_{1,i}$ are error terms with unknown distributions, mean independent of $Z_i, \boldsymbol{X}_i$ by definition — but possibly correlated with D_i. Z_i is conditionally independent of potential outcomes, so that $U_{0,i}, U_{1,i} \perp\!\!\!\perp Z_i$.

The first-stage regression of $Z \to Y$ has unbiased estimates, since $Z_i \perp\!\!\!\perp D_i(.) \big| \boldsymbol{X}_i$. Put $\pi(z'; \boldsymbol{X}) = \mathbb{E}\left[D_i(z') \,|\, \boldsymbol{X}\right]$, and $\eta_{z',i} = D_i(z') - \pi(z'; \boldsymbol{X})$ the first-stage error terms.

$$\begin{aligned}
D_i &= Z_i D_i(1) + (1 - Z_i) D_i(0) \\
&= D_i(0) + Z_i \left[D_i(1) - D_i(0)\right] \\
&= \underbrace{\pi(0; \boldsymbol{X}_i)}_{\text{Intercept, } := \phi + \zeta(\boldsymbol{X}_i)} + \underbrace{Z_i \mathbb{E}\left[\pi(1; \boldsymbol{X}_i) - \pi(0; \boldsymbol{X}_i)\right]}_{\text{Regressor, } := \overline{\pi} Z_i} + \underbrace{(1 - Z_i)\eta_{0,i} + Z_i \eta_{1,i}}_{\text{Errors, } := \eta_i} \\
\implies \mathbb{E}\left[D_i \,|\, Z_i, \boldsymbol{X}_i\right] &= \phi + \overline{\pi} Z_i + \zeta(\boldsymbol{X}_i).
\end{aligned}$$

Since the ignorability assumption gives $\mathbb{E}\left[Z_i \eta_{z',i}\right] = \mathbb{E}\left[Z_i\right] \mathbb{E}\left[\eta_{z',i}\right] = 0$, for each $z' = 0, 1$.

By the same argument Z_i is also assumed independent of potential outcomes $Y_i(., .)$, so that $U_{0,i}, U_{1,i} \perp\!\!\!\perp Z_i$. Thus, the reduced form regression $Z \to Y$ also leads to unbiased estimates for the ATE.

The same cannot be said of the regression that estimates direct and indirect effects, without further assumptions.

$$\begin{aligned}
Y_i &= Z_i Y_i(1, D_i(1)) + (1 - Z_i) Y_i(0, D_i(0)) \\
&= Z_i D_i Y_i(1, 1) \\
&\quad + (1 - Z_i) D_i Y_i(0, 1) \\
&\quad + Z_i (1 - D_i) Y_i(1, 0) \\
&\quad + (1 - Z_i)(1 - D_i) Y_i(0, 0) \\
&= Y_i(0, 0) \\
&\quad + Z_i \left[Y_i(1, 0) - Y_i(0, 0)\right] \\
&\quad + D_i \left[Y_i(0, 1) - Y_i(0, 0)\right] \\
&\quad + Z_i D_i \left[Y_i(1, 1) - Y_i(1, 0) - (Y_i(0, 1) - Y_i(0, 0))\right]
\end{aligned}$$

And so Y_i can be written as a regression equation in terms of the observed factors and error terms.

$$\begin{aligned}
Y_i = & \mu_0(0; \boldsymbol{X}_i) \\
& + D_i \left[\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\right] \\
& + Z_i \left[\mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\right] \\
& + Z_i D_i \left[\mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - (\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i))\right] \\
& + U_{0,i} + D_i \left(U_{1,i} - U_{0,i}\right) \\
= & \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) + (1 - D_i) U_{0,i} + D_i U_{1,i}
\end{aligned}$$

With the following definitions:

(a) $\alpha = \mathbb{E}\left[\mu_0(0; \boldsymbol{X}_i)\right]$ and $\varphi(\boldsymbol{X}_i) = \mu_0(0; \boldsymbol{X}_i) - \alpha$ are the intercept terms.

(b) $\beta = \mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)$ is the indirect effect under $Z_i = 0$

(c) $\gamma = \mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)$ is the direct effect under $D_i = 0$.

(d) $\delta = \mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - (\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i))$ is the interaction effect.

(e) $(1 - D_i) U_{0,i} + D_i U_{1,i}$ is the remaining error term.

This sequence gives us the resulting regression equation:

$$\begin{aligned}
\mathbb{E}\left[Y_i \mid Z_i, D_i, \boldsymbol{X}_i\right] \;=\; & \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\
& + (1 - D_i)\, \mathbb{E}\left[U_{0,i} \mid D_i = 0, \boldsymbol{X}_i\right] + D_i \mathbb{E}\left[U_{1,i} \mid D_i = 1, \boldsymbol{X}_i\right]
\end{aligned}$$

Taking the conditional expectation, and collecting for the expressions of the direct and indirect effects:

$$\begin{aligned}
\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] &= \mathbb{E}\left[\gamma + \delta D_i\right] \\
\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] &= \mathbb{E}\left[\overline{\pi}\left(\beta + Z_i \delta + \widetilde{U}_i\right)\right]
\end{aligned}$$

These equations have simpler expressions after assuming constant treatment effects in a linear framework; I have avoided this as having compliers, and controlling for observed factors $\boldsymbol{X}_i$ only makes sense in the case of heterogeneous treatment effects.

These terms are conventionally estimated in a simultaneous regression (Imai et al. 2010). If sequential ignorability does not hold, then the regression estimates from estimating the mediation equations (without adjusting for the contaminated bias term) suffer from omitted variables bias.

$$\mathbb{E}_{\boldsymbol{X}_i}\left[\mathbb{E}\left[Y_i \,|\, Z_i = D_i = 0, \boldsymbol{X}_i\right]\right] = \mathbb{E}\left[\alpha\right] + \mathbb{E}\left[U_{0,i} \,|\, D_i = 0\right]$$

$$\mathbb{E}_{\boldsymbol{X}_i}\left[\mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 0, \boldsymbol{X}_i\right]\right] = \mathbb{E}\left[\beta\right] + \left(\mathbb{E}\left[U_{1,i} \,|\, D_i = 1\right] - \mathbb{E}\left[U_{0,i} \,|\, D_i = 0\right]\right)$$

$$\mathbb{E}_{\boldsymbol{X}_i}\left[\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i = 0, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 0, \boldsymbol{X}_i\right]\right] = \mathbb{E}\left[\gamma\right] + \mathbb{E}\left[U_{0,i} \,|\, D_i = 0\right]$$

$$\mathbb{E}_{\boldsymbol{X}_i}\begin{bmatrix}\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i = 0, \boldsymbol{X}_i\right] \\ -\left(\mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 0, \boldsymbol{X}_i\right]\right)\end{bmatrix} = \mathbb{E}\left[\delta\right]$$

And so the ADE and AIE estimates are contaminated by these bias terms. Additionally, the AIE estimates refers to gains from the mediator among $D(z)$ compliers (not the entire average), so will be biased when not accounting for $\widetilde{U}_i$, too.

A.4 Roy Model and Sequential Ignorability

Proof of Proposition 1.

Suppose Z_i is ignorable, and selection-into-D_i follows a Roy model, with the definitions in Section 2. If selection-into-D_i is degenerate on $U_{0,i}, U_{1,i}$:

$$\mathbb{E}\left[D_i \,|\, Z_i, \boldsymbol{X}_i, U_{1,i} - U_{0,i} = u\right] = \mathbb{E}\left[D_i \,|\, Z_i, \boldsymbol{X}_i, U_{1,i} - U_{0,i} = u'\right], \text{ for all } u, u' \text{ in the range of } U_{1,i} - U_{0,i}.$$

In this case, the control set $\boldsymbol{X}_i$ and the costs $\mu_c, U_{c,i}$ are the only determinants of selection-into-D_i — and, $U_{0,i}, U_{1,i}$ play no role. This could be achieved by either assuming that unobserved gains are degenerate (the researcher had observed everything in $\boldsymbol{X}_i$), or selection-into-D_i had been disrupted in some fashion (e.g., by a natural experiment design for D_i).

To motivate a contraposition argument, suppose D_i is ignorable conditional on $Z_i, \boldsymbol{X}_i$. For each $z', d' = 0, 1$

$$\begin{aligned}
& D_i \perp\!\!\!\perp Y_i(z', d') \;\mid\; \boldsymbol{X}_i, Z_i = z' \\
\implies & D_i \perp\!\!\!\perp \mu_{d'}(z'; \boldsymbol{X}_i) + U_{d',i} \;\mid\; \boldsymbol{X}_i, Z_i = z' \\
\implies & D_i \perp\!\!\!\perp U_{d',i} \;\mid\; \boldsymbol{X}_i, Z_i = z' \\
\implies & D_i \perp\!\!\!\perp U_{1,i} - U_{0,i} \;\mid\; \boldsymbol{X}_i, Z_i = z' \\
\implies & \mathbb{E}\left[D_i \,|\, U_{1,i} - U_{0,i} = u', \boldsymbol{X}_i, Z_i = z'\right] = \mathbb{E}\left[D_i \,|\, \boldsymbol{X}_i, Z_i = z'\right] \\
& \text{for all } u' \text{ in the range of } U_{1,i} - U_{0,i}.
\end{aligned}$$

This final implication is that selection-into-D_i is degenerate on $U_{0,i}, U_{1,i}$. Thus, a contraposition argument has that if selection-into-D_i is non-degenerate on $U_{0,i}, U_{1,i}$, then D_i is not ignorable.

A.5 Monotonicity $\implies$ Selection Model, in a CM Setting.

Proof that (conditional) monotonicity implies a selection model representation in a CM setting. This proof is an applied example of the Vytlacil (2002) equivalence result, now including conditioning covariates $\boldsymbol{X}_i$, and is presented merely as a validation exercise.

Assume condition monotonicity CF–1 holds, for any treatment values $z < z'$ and any covariate value $\boldsymbol{X}_i = \boldsymbol{x}$.

$$\Pr\left(D_i(z') \geq D_i(z) \,|\, \boldsymbol{x}\right) = 1.$$

For each value of $\boldsymbol{X}_i = \boldsymbol{x}$ and any treatment values $z < z'$, we first define:

- $\mathcal{A} = \{i : D_i(z) = D_i(z') = 1\}$, always-mediators
- $\mathcal{N} = \{i : D_i(z) = D_i(z') = 0\}$, never-mediators
- $\mathcal{C} = \{i : D_i(z) = 0, D_i(z') = 1\}$, mediator-compliers.

For any mediator complier $i \in \mathcal{C}$, partition the set as follows.

- $\mathcal{Z}_1(i) = \{z' : D_i(z') = 1\}$, treatment values where i takes the mediator
- $\mathcal{Z}_0(i) = \{z' : D_i(z') = 0\}$, treatment values where i doesn't take the mediator.

Note that having binary $Z_i = 0, 1$ reduces this to the simple case of $\mathcal{Z}_0(i) = \{0\}$, and $\mathcal{Z}_1(i) = \{1\}$. The equivalence result holds for continuous values of Z_i, so continue with the more general $\mathcal{Z}_0(i), \mathcal{Z}_1(i)$ notation.

By monotonicity, we have

$$\sup_{z' \in \mathcal{Z}_0(i)} \pi(z'; \boldsymbol{x}) \leq \inf_{z' \in \mathcal{Z}_1(i)} \pi(z'; \boldsymbol{x}), \quad \text{for any } i \in \mathcal{C}$$

where $\pi(z'; \boldsymbol{x}) = \Pr\left(D_i = 1 \,|\, Z_i = z', \boldsymbol{X}_i = \boldsymbol{x}\right)$ is the mediator propensity score. A simple proof by contradiction verifies this statement (Vytlacil 2002, Lemma 1).

Now we construct V_i as follows:

$$V_i = \begin{cases} 1, & \text{if } i \in \mathcal{N} \\ 0, & \text{if } i \in \mathcal{A} \\ \inf_{z' \in \mathcal{Z}_1(i)} \pi(z'; \boldsymbol{x}), & \text{if } i \in \mathcal{C}. \end{cases}$$

Define $\psi(z'; \boldsymbol{x}) = \pi(z'; \boldsymbol{x})$. Then we can represent $D_i(z')$ as a selection model,

$$D_i(z') = \mathbb{1}\left\{\psi(z'; \boldsymbol{X}_i) \geq V_i\right\}, \quad \text{for } z' = 0, 1.$$

We can verify this works:

- For $i \in \mathcal{A}$: $V_i = 0$ and $\psi(z'; \boldsymbol{x}) \geq 0$ for all z', so $D_i(z') = 1$
- For $i \in \mathcal{N}$: $V_i = 1$ and $\psi(z'; \boldsymbol{x}) \leq 1$ for all z', with $\psi(z'; \boldsymbol{x}) < 1$ for $z' \in \mathcal{Z}_0(i)$, so $D_i(z') = 0$ for $z' \in \mathcal{Z}_0(i)$
- For $i \in \mathcal{C}$: $V_i = \inf_{z' \in \mathcal{Z}_1(i)} \pi(z'; \boldsymbol{x})$
  - When $z' \in \mathcal{Z}_1(i)$: $\psi(z'; \boldsymbol{x}) \geq \inf_{z'' \in \mathcal{Z}_1(i)} \pi(z''; \boldsymbol{x}) = V_i$, so $D_i(z') = 1$
  - When $z' \in \mathcal{Z}_0(i)$: $\psi(z'; \boldsymbol{x}) < \inf_{z'' \in \mathcal{Z}_1(i)} \pi(z''; \boldsymbol{x}) = V_i$, so $D_i(z') = 0$.

Therefore, the construction $D_i(z') = \mathbb{1}\{\psi(z'; \boldsymbol{X}_i) \geq V_i\}$ is a valid representation of the selection process under monotonicity.

This selection model can be transformed to one with a uniform distribution, to get the general selection model of Heckman & Vytlacil (2005). Let $F_V\big(. \,\big|\, \boldsymbol{X}_i\big)$ be the conditional cumulative density function of V_i given $\boldsymbol{X}_i$. Define

$$\begin{aligned} U_i &= F_V\left(V_i \mid \boldsymbol{X}_i\right) \\ \pi(z'; \boldsymbol{X}_i) &= F_V\left(\psi(z'; \boldsymbol{X}_i) \mid \boldsymbol{X}_i\right) = \Pr\left(D_i = 1 \,|\, Z_i = z', \boldsymbol{X}_i\right) \end{aligned}$$

We can then equivalently represent the mediator choice as the transformed selection model

$$D_i(z') = \mathbb{1}\{\pi(z'; \boldsymbol{X}_i) \geq U_i\}, \quad \text{for } z' = 0, 1$$

where $U_i \,|\, \boldsymbol{X}_i \sim \text{Uniform}(0, 1)$ by the probability integral transformation.

A.6 Control Function (CF) Identification of the Second-stage

Proof of Proposition 2. This proof relies heavily on the notation and reasoning of Kline & Walters (2019) for an IV setting.

By Assumption CF–1 (mediator monotonicity), selection-into-mediator can be represented as a threshold-crossing selection model.

$$D_i(z') = \mathbb{1}\{\pi(z'; \boldsymbol{X}_i) \geq U_i\}, \text{ for } z' = 0, 1$$

where $U_i = F_V\left(V_i \mid \boldsymbol{X}_i\right)$ follows a uniform distribution on $[0, 1]$, and $\pi(z'; \boldsymbol{X}_i) = \mathbb{E}\left[D_i \,|\, Z_i = z', \boldsymbol{X}_i\right]$ is the mediator propensity score.

The threshold crossing selection model represents individuals who refuse the mediator as follows:

$$D_i = 0 \implies \pi(Z_i; \boldsymbol{X}_i) < U_i$$

Our objective is to determine $\mathbb{E}\left[U_{0,i} \,|\, D_i = 0, Z_i, \boldsymbol{X}_i\right]$, which can then be written as

$$\mathbb{E}\left[U_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i, Z_i, \boldsymbol{X}_i\right].$$

Since Z_i is ignorable, we have:

$$\mathbb{E}\left[U_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i, Z_i, \boldsymbol{X}_i\right] = \mathbb{E}\left[U_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right]$$

Assumption CF–2 has $\text{Cov}(U_i, U_{0,i}) \neq 0$. This non-zero covariance implies statistical dependence between the selection error and outcome error. This dependence allows us to represent $U_{0,i}$ using a linear projection. We use $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$ rather than U_i directly in the projection to allow for flexibility in how the selection error affects outcomes. The linear projection can be written as follows

$$U_{0,i} = \rho_0\big(F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V\big) + \varepsilon_{0,i},$$

where

- $\mu_V = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)\right]$ is the mean of $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$
- $\rho_0 = \frac{\text{Cov}\left(U_{0,i}, F_V^{-1}(U_i|\boldsymbol{X}_i)\right)}{\text{Var}\left(F_V^{-1}(U_i|\boldsymbol{X}_i)\right)}$ is the projection coefficient
- $\varepsilon_{0,i}$ is a residual with $\mathbb{E}\left[\varepsilon_{0,i} \,\middle|\, F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)\right] = 0$.

The coefficient ρ_0 is the slope in the best linear predictor of $U_{0,i}$ given $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$, and is chosen to ensure that the residual $\varepsilon_{0,i}$ is uncorrelated with $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$. This property is critical for our identification strategy, as it allows us to isolate the part of $U_{0,i}$ that is related to selection-into-D_i.

The non-zero covariance condition in CF ensures $\rho_0 \neq 0$, so is relevant. Since U_i and $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$ are related by a monotonic transformation (the inverse cumulative density function), the covariance $\text{Cov}(U_i, U_{0,i}) \neq 0$ implies $\text{Cov}(F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right), U_{0,i}) \neq 0$.

Given the linear projection of $U_{0,i}$ onto $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$, we can compute the conditional expectation:

$$\mathbb{E}\left[U_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right] = \mathbb{E}\left[\rho_0\big(F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V\big) + \varepsilon_{0,i} \,\middle|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right]$$

Since $\mathbb{E}\left[\varepsilon_{0,i} \,\middle|\, F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)\right] = 0$ by construction, and U_i is a function of $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$, we have

$$\mathbb{E}\left[\varepsilon_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right] = 0.$$

Therefore:

$$\mathbb{E}\left[U_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right] = \rho_0 \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right].$$

This gives us the control function representation:

$$\mathbb{E}\left[U_{0,i} \,|\, D_i = 0, Z_i, \boldsymbol{X}_i\right] = \rho_0 \lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big)$$

where $\lambda_0(p') = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, p' < U_i\right]$. The control function $\lambda_0(p')$ captures the expected value of the transformed selection term, conditional on being above the threshold $p' \in (0, 1)$.

The same sequence of steps for mediator takers, $D_i = 1$, gives the other CF:

$$\mathbb{E}\left[U_{1,i} \,|\, D_i = 1, Z_i, \boldsymbol{X}_i\right] = \rho_1 \lambda_1\big(\pi(Z_i; \boldsymbol{X}_i)\big),$$

where $\lambda_1(p') = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, U_i \leq p'\right]$ for $p' \in (0, 1)$.

The relationship between $\lambda_0(p')$ and $\lambda_1(p')$ can be derived as:

$$\lambda_1(p') = -\lambda_0(p')\left(\frac{1-p'}{p'}\right), \text{ for } p' \in (0, 1).$$

This relationship ensures consistency in the CF approach across both groups (Kline & Walters 2019).

Assumption CF–3 (mediator take-up cost instrument $\boldsymbol{X}_i^{\text{IV}}$) ensures identification of the propensity score function $\pi(z'; \boldsymbol{X}_i)$ in the first stage by providing valid instrumental variation. This variation allows us to identify the propensity score, and consequently the control functions λ_0 and λ_1.

Combining all elements, the conditional expectation of Y_i given $Z_i, D_i, \boldsymbol{X}_i$ is

$$\begin{aligned}\mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i\right] &= \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\ &\quad + (1 - D_i)\, \mathbb{E}\left[U_{0,i} \,|\, D_i = 0\right] + D_i \mathbb{E}\left[U_{1,i} \,|\, D_i = 1\right].\end{aligned}$$

Substitute the CFs, as follows:

$$(1-D_i)\mathbb{E}\left[U_{0,i} \,|\, Z_i, D_i, \boldsymbol{X}_i\right] + D_i \mathbb{E}\left[U_{1,i} \,|\, Z_i, D_i, \boldsymbol{X}_i\right] = (1-D_i)\rho_0\lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big) + D_i\rho_1\lambda_1\big(\pi(Z_i; \boldsymbol{X}_i)\big).$$

This gives the final result,

$$\begin{aligned}\mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i\right] &= \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\ &\quad + \rho_0 (1 - D_i)\, \lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big) + \rho_1 D_i \lambda_1\big(\pi(Z_i; \boldsymbol{X}_i)\big).\end{aligned}$$

All parameters — $\alpha, \beta, \gamma, \delta, \varphi(.), \rho_0, \rho_1$ — are identified once we control for selection bias through the CFs λ_0, λ_1, with $\pi(z'; \boldsymbol{X}_i)$ identified separately in the first-stage. λ_0, λ_1 can be assumed to be certain functions (say, the inverse Mills ratio in Heckman 1979), or treated as non-parametric parameters to be estimated with the moment equality $\lambda_1(p') = -\lambda_0(p')\left(\frac{1-p'}{p'}\right)$ for $p \in (0, 1)$.

A.7 Control Function (CF) Identification of the ADE and AIE

Proof of Theorem CF.

Assume CF–1, CF–2, CF–3 hold. Then Proposition 2 has $\alpha, \beta, \gamma, \delta, \varphi(.), \rho_0, \rho_1$ identified in a regression. The following composes the ADE and AIE from these parameters.

For the ADE,

$$
\begin{aligned}
\mathbb{E}[\gamma + \delta D_i] &= \mathbb{E}\left[\Big(\mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\Big) + D_i\Big(\mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - \big(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\big)\Big)\right] \\
&= \mathbb{E}\left[D_i\Big(\mu_1(1; \boldsymbol{X}_i) - \mu_1(0; \boldsymbol{X}_i)\Big) + (1 - D_i)\Big(\mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\Big)\right] \\
&= \mathbb{E}\left[D_i\Big(Y_i(1, 1) - U_{1,i} - \big(Y_i(0, 1) - U_{1,i}\big)\Big) + (1 - D_i)\Big(Y_i(1, 0) - U_{0,i} - \big(Y_i(0, 0) - U_{0,i}\big)\Big)\right] \\
&= \mathbb{E}\left[D_i\Big(Y_i(1, 1) - Y_i(0, 1)\Big) + (1 - D_i)\Big(Y_i(1, 0) - Y_i(0, 0)\Big)\right] \\
&= \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] \\
&= \text{ADE}.
\end{aligned}
$$

Identification is similar for the AIE, but also involves the complier adjustment term.

$$
\begin{aligned}
(\rho_1 - \rho_0)\,\Gamma\big(\pi(0; \boldsymbol{X}_i),\, \pi(1; \boldsymbol{X}_i)\big) &= (\rho_1 - \rho_0)\,\frac{\pi(1; \boldsymbol{X}_i)\lambda_1(\pi(1; \boldsymbol{X}_i)) - \pi(0; \boldsymbol{X}_i)\lambda_1(\pi(0; \boldsymbol{X}_i))}{\pi(1; \boldsymbol{X}_i) - \pi(0; \boldsymbol{X}_i)} \\
&= (\rho_1 - \rho_0)\,\mathbb{E}\left[F_V^{-1}(U_i | \boldsymbol{X}_i) - \mu_V \,\middle|\, \pi(0; \boldsymbol{X}_i) < U_i \leq \pi(1; \boldsymbol{X}_i), \boldsymbol{X}_i\right] \\
&= (\rho_1 - \rho_0)\,\mathbb{E}\left[F_V^{-1}(U_i | \boldsymbol{X}_i) - \mu_V \,\middle|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right] \\
&= \mathbb{E}\left[\rho_1\big(F_V^{-1}(U_i | \boldsymbol{X}_i) - \mu_V\big) \,\middle|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right] \\
&\quad - \mathbb{E}\left[\rho_0\big(F_V^{-1}(U_i | \boldsymbol{X}_i) - \mu_V\big) \,\middle|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right] \\
&= \mathbb{E}\left[U_{1,i} - U_{0,i} \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right].
\end{aligned}
$$

This complier adjustment was first presented for an IV setting by Kline & Walters (2019).

Collecting for the AIE,

$$
\begin{aligned}
&\mathbb{E}\left[\overline{\pi}\left(\beta + \delta Z_i + (\rho_1 - \rho_0)\Gamma\left(\pi(0; \boldsymbol{X}_i),\, \pi(1; \boldsymbol{X}_i)\right)\right)\right] \\
&= \mathbb{E}\left[\overline{\pi}\left(\left(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\right) + Z_i\left(\mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - \left(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\right)\right)\right)\right] \\
&\quad + \mathbb{E}\left[\overline{\pi}\, \mathbb{E}\left[U_{1,i} - U_{0,i} \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[\overline{\pi}\left(Z_i\left(\mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i)\right) + (1 - Z_i)\left(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\right)\right)\right] \\
&\quad + \mathbb{E}\left[\overline{\pi}\, \mathbb{E}\left[U_{1,i} - U_{0,i} \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[\overline{\pi}\left(\mu_1(Z_i, \boldsymbol{X}_i) - \mu_0(Z_i, \boldsymbol{X}_i) + \mathbb{E}\left[U_{1,i} - U_{0,i} \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right]\right)\right] \\
&= \mathbb{E}\left[\overline{\pi}\, \mathbb{E}\left[\mu_1(Z_i, \boldsymbol{X}_i) - \mu_0(Z_i, \boldsymbol{X}_i) + U_{1,i} - U_{0,i} \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[D_i(1) - D_i(0) \,|\, \boldsymbol{X}_i\right]\; \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \,|\, \boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] \\
&= \text{AIE}.
\end{aligned}
$$

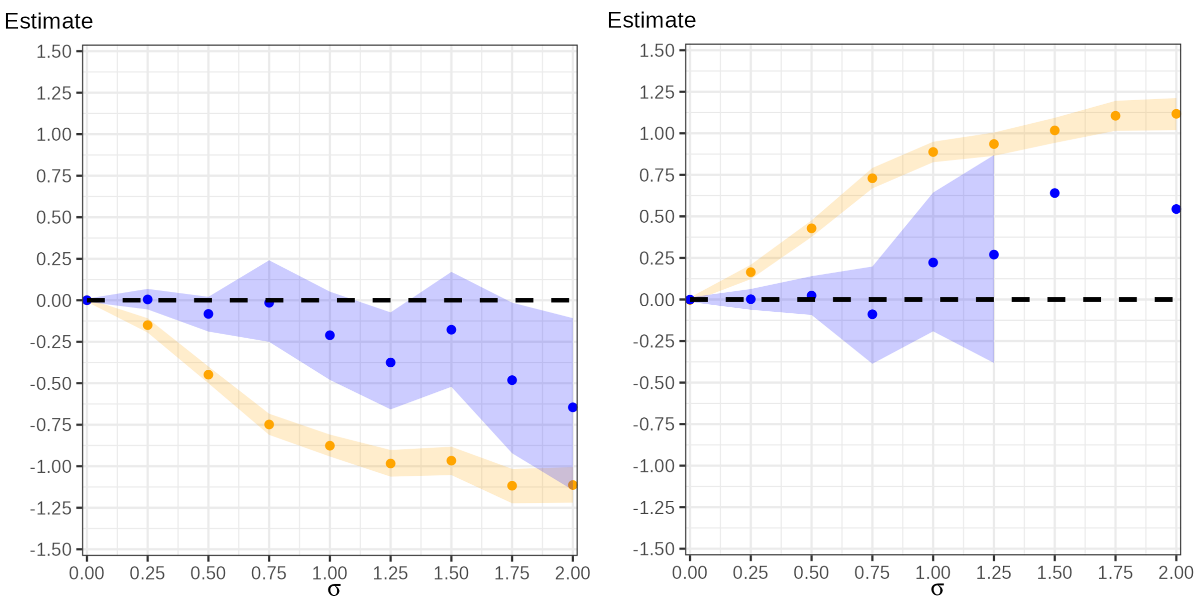
A.8 Control Function Simulation

A number of statistical packages, for the R language (R Core Team 2023), made the simulation analysis for this paper possible.

- *Tidyverse* (Wickham, Averick, Bryan, Chang, McGowan, François, Grolemund, Hayes, Henry, Hester, Kuhn, Pedersen, Miller, Bache, Müller, Ooms, Robinson, Seidel, Spinu, Takahashi, Vaughan, Wilke, Woo & Yutani 2019) collected tools for data analysis in the R language.
- *Splines* (Wang & Yan 2021) allows semi-parametric estimation, using splines, in the R language.
- *Mediate* (Tingley, Yamamoto, Hirose, Keele & Imai 2014) automates the sequential-ignorability estimates of CM effects (Imai et al. 2010) in the R language.

Figure A1: Point Estimates of CM Effects, OLS and Control Function versus True Value.

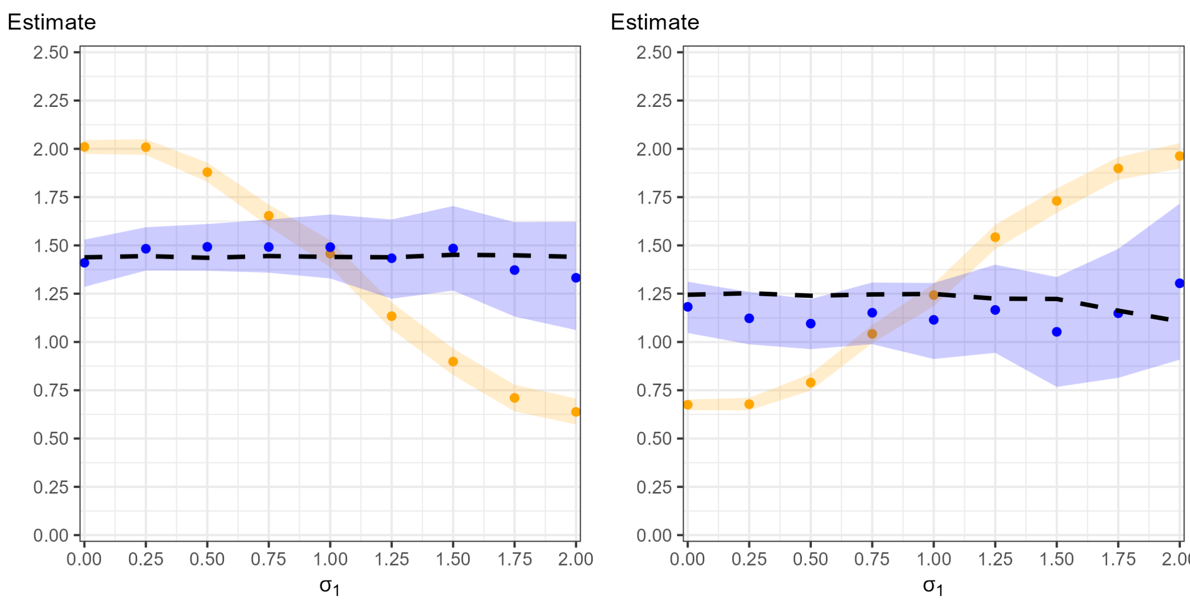
(a) $\widehat{\text{ADE}} - \text{ADE}$. **(b)** $\widehat{\text{AIE}} - \text{AIE}$.



Note: These figures show the OLS and control function point estimates of the ADE and AIE, for $N = 10,000$ sample size, minus the true value of the ADE and AIE, respectively. y-axis value of zero means the point estimate had estimated the ADE, or AIE, exactly. Points are points estimates from data simulated with a given $\rho = 0.5$ value, varying the $\sigma_0 = \sigma, \sigma_1 = 2\sigma$ values. Orange represents OLS estimates, blue the control function approach. Shaded regions are the 95% confidence intervals from 1,000 bootstraps each.

Figure A2: OLS versus Control Function Estimates of CM Effects, varying σ_1 relative to $\sigma_0 = 1$.

(a) ADE. **(b)** AIE.



Note: These figures show the OLS and control function estimates of the ADE and AIE, for $N = 10,000$ sample size. The black dashed line is the true value, points are points estimates from data simulated with a given $\rho = 0.5, \sigma_0 = 1$ and σ_1 varied across $[0, 2]$. Shaded regions are the 95% confidence intervals; orange are the OLS estimates, blue the control function approach.