Causal Mediation in Natural Experiments

Senan Hogan-Hennessy*
Economics Department, Cornell University†

First draft: 12 February 2025
This version: 17 June 2025

Working Paper, 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 randomisation. 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 quasi-random assignment of the initial treatment. Individuals' choice to take (or refuse) a mediator based on costs and benefits is inconsistent with this assumption, suggesting in-practice estimates of causal mechanisms have no causal interpretation. I consider an alternative approach to credibly estimate CM effects, using control function methods and relying on instrumental variation in mediator take-up costs. Simulations confirm this approach corrects for bias in conventional CM estimates, providing parametric and semi-parametric methods. This approach gives applied researchers an alternative method to estimate CM effects when an initial treatment is quasi-randomly assigned, but the mediator is not, 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 Neil Cholli, Hyewon Kim, Jiwoo Kim, Lukáš Laffers, Jiwon Lee, Douglas Miller, Zhuan Pei, Brenda Prallon, and Evan Riehl. 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, https://github.com/shoganhennessy/mediation-natural-experiment.

†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 health insurance causally improve health outcomes (**?**)? Natural experiments are settings which answer 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, do causal gains from health insurance come mostly from starting to utilise healthcare more often, or are there other direct effects? This study of mechanisms behind causal effects broadens the economic understanding of social settings 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. For example, if individuals had been choosing to seek medical care more frequently with new health insurance, then underlying health conditions would confound estimates of the direct and indirect effects of health insurance through using more healthcare. This approach only leads to credible causal estimates if the mediator is also quasi-randomly assigned. 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 condition 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 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 costs 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 conventional approach to CM assumes that the original treatment, and the subsequent mediator, are both ignorable (**?**). This approach arose in the statistics literature, and is widely used in social sciences to estimate CM effects in observational studies. Informal mechanisms analyses in applied economics allude to CM methods (despite masquerading under an alternative moniker), and so unintentionally import this identifying assumption.

Assuming the mediator is ignorable (i.e., quasi-randomly assigned or satisfies selection-on-observables) conveniently sidesteps the consideration of individual choice by assuming that either people made decisions to take/refuse a mediator naïvely, or a researcher controlled for 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 controlling for a proposed mediator (**?**). 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 (**??**), functional form restrictions (**?**), partial identification (**?**), or a hypothesis test of full mediation through observed channels (**?**) — see **?** 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 **?** model as a benchmark for judging the **?** mediator ignorability assumption in a natural experiment setting, and find it unlikely to hold in a natural experiment setting.[1] This motivates a solution to the identification problem inspired by classic labour economic work, which also uses the Roy model as a benchmark (**??**). 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 mediator take-up respond only positively to the initial treatment (monotonicity), which implies mediator selection follows a selection model. Second, it assumes that mediator take-up is motivated by mediator benefits. Last, it requires a valid instrument for mediator take-up, to avoid relying on parametric assumptions on unobserved selection. This approach to identifying CM effects (despite selection-into-mediator) imports insights from the instrumental variables literature, connecting CM to the marginal treatment effects literature (**?????**).[2] Using a control function to estimate CM effects builds on the influential **?** approach, marrying the CM literature with labour economic theory on selection-into-treatment for the first time.

This paper proceeds as follows. **??** introduces the formal framework for CM, and develops expressions for bias in CM estimates in natural experiments. **??** describes this bias in applied settings with (1) a regression framework, (2) a setting with selection based on costs and benefits. **??** shows how a control function can effectively purge this bias from CM estimates. **??** demonstrates how to estimate CM effects with this approach, with either parametric or semi-parametric methods, giving supporting simulation evidence. **??** concludes.

1 Average Direct and Indirect Effects

CM decomposes causal effects into two channels, through a mediator (indirect effect) and through all other paths (direct effect). To develop notation, write $Z_i = 0, 1$ for a binary treatment, $D_i = 0, 1$ a binary mediator, and Y_i a continuous outcome.[3] D_i, Y_i are a sum of

[1] An alternative method to estimate CM effects is ensuring treatment and mediator ignorability holds by a running two randomised controlled trials (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 (**??**) and in practice (**??**).

[2] Indeed, this paper does not invent control function methods, instead noting their applicability in this setting. See **??** for general overviews of the approach, and **?** for a CM setting with two instruments that notes the connection to control function methods.

[3] This paper exclusively focuses on the binary case. See **?** for a discussion of CM with continuous treatment and/or mediator, and the assumptions required.

their potential outcomes,

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

Assume treatment 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_i on D_i:

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

It is common in the economics literature to assume that Z_i influences D_i in at most one direction, $\Pr\left(D_i(0) \leq D_i(1)\right) = 1$ — monotonicity (**?**). 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_i on Y_i, 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_i affects outcome Y_i directly, and indirectly via the $D_i(Z_i)$ channel, with no reverse causality. **??** visualises the design, where the direction arrows denote the causal direction. CM aims to decompose the ATE of Z_i on Y_i into these two separate pathways:

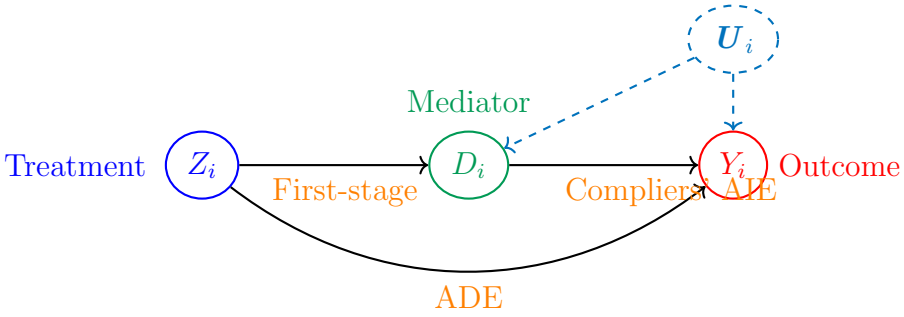
$$\text{Average Direct Effect (ADE):} \quad \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right],$$
$$\text{Average Indirect Effect (AIE):} \quad \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right].$$

Estimating the AIE answers the following question: how much of the causal effect Z_i on Y_i goes through the D_i channel? If a researcher is studying the health gains of health insurance (**?**), 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_i channel?[5] For the health

[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.

[5] In a non-parametric setting it is not necessary that ADE + AIE = ATE. See **?** for this point in full.

Figure 1: Structural Causal Model for Causal Mediation.



Note: This figure shows the structural causal model behind causal mediation, where arrows represent causal effects — e.g., $Z_i \to D_i$ means Z_i affects D_i with no reverse causality. Compliers' AIE refers to the AIE local to $D_i(Z_i)$ compliers, so that AIE = average first-stage × compliers' AIE. $\boldsymbol{U}_i$ represents this paper's focus on the case that D_i is not ignorable by showing an unobserved confounder. **??** defines $\boldsymbol{U}_i$ in an applied setting.

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 Identification of 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 $\boldsymbol{X}_i$.

Definition 1. *Sequential Ignorability (**?**)*

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

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

Sequential ignorability assumes that the initial treatment Z_i is ignorable conditional on $\boldsymbol{X}_i$ (as has already been assumed above). It then also assumes that, after Z_i is assigned, that D_i is ignorable conditional on $\boldsymbol{X}, Z_i$ (hereafter, mediator ignorability). If **??**(**??**) and **??**(**??**) hold, then the ADE and AIE are identified by two-stage mean differences conditioning on $\boldsymbol{X}_i$.[6]

[6]In addition, a common support condition for both Z_i, D_i (across $\boldsymbol{X}_i$) is necessary. **?** show a general identification statement; I show identification in terms of two-stage regression, notation for which is more familiar in economics. Appendix **??** states the **?** identification result, and then develops the two-stage

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

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

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 (**?**). 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., **??**). This section focuses on problems that plague CM by selection-on-observables, regardless of estimation method.

1.2 Non-identification of Causal Mediation (CM) Effects

Applied researchers often use a natural experiment to study settings where treatment Z_i is ignorable, justifying assumption **??**(**??**). Rarely do they also have access to an additional, overlapping natural experiment to isolate random variation in D_i — to justify mediator ignorability **??**(**??**). One might consider conventional CM methods in such a setting to learn about the mechanisms behind the causal effect Z_i on Y_i. 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. If* **??**(**??**) *holds, and* **??**(**??**) *does not, then CM estimands are contaminated by selection bias and group differences. Proof: see Appendix* **??**.

Below I present the relevant selection bias and group difference terms, omitting the conditional on $\boldsymbol{X}_i$ notation for brevity.

regression notation which holds as a consequence of sequential ignorability.

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 selection bias terms would equal zero if the mediator had been ignorable **??**(**??**), 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 (**?**). 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 **?**, or M-bias in **?**).

The AIE group differences term is longer, because the indirect effect is comprised of the

[7] The bias terms here mirror those in **??** for a single D_i on Y_i treatment effect, when D_i is not ignorable: $\mathbb{E}\left[Y_i \mid D_i = 1\right] - \mathbb{E}\left[Y_i \mid D_i = 0\right] = \text{ATE} + \underbrace{\Big(\mathbb{E}\left[Y_i(., 0) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(., 0) \mid D_i = 0\right]\Big)}_{\text{Selection Bias}} + \underbrace{\Pr\left(D_i = 0\right)\left(\text{ATT} - \text{ATU}\right)}_{\text{Group-differences Bias}}$.

effect of D_i local to $D_i(Z_i)$ compliers.

$$\begin{aligned}\text{AIE} &= \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] \\ &= \mathbb{E}\left[D_i(1) - D_i(0)\right] \underbrace{\mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \mid D_i(0) = 0, D_i(1) = 1\right]}_{\text{Average } D_i \text{ on } Y_i \text{ effect among } D_i(Z_i) \text{ compliers}}\end{aligned}$$

It is important to acknowledge the mediator compliers here, because the AIE is the treatment effect going through the $D_i(Z_i)$ 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_i on Y_i, then this is losing focus on the definition of the AIE; it is not about the causal effect D_i on Y_i, it is about the causal effect $D_i(Z_i)$ on Y_i.

The group difference bias term 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. This distinction between average effects and complier average effects in the AIE is skipped over by the "controlled effect" definitions of **?**.

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 (**?**). 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 **??** 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 **?**. For each $z', d' = 0, 1$, put $\mu_{d'}(z'; \boldsymbol{X}_i) = \mathbb{E}\left[Y_i(z', d') \mid \boldsymbol{X}_i\right]$ and the corresponding error terms, $U_{d',i} = Y_i(z', d') - \mu_{d'}(z'; \boldsymbol{X}_i)$, so we have the following expressions:

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

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

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

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

This is not consequence of linearity assumptions; the regression functions allow for unconstrained heterogenous treatment effects. This is because Z_i, D_i are categorical, and if either were instead continuously distributed then this representative would not necessarily hold true. First-stage (**??**) is identified, with $\theta + \zeta(\boldsymbol{X}_i)$ the intercept, and $\overline{\pi}$ the first-stage average compliance rate (conditional on $\boldsymbol{X}_i$). Second-stage (**??**) has the following definitions, and is not identified thanks to omitted variables bias. See Appendix **??** 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 conditional on $Z_i = 0, \boldsymbol{X}_i$.

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

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

(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.

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

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

By construction, $\boldsymbol{U}_i \coloneqq (U_{0,i}, U_{1,i})$ is an unobserved confounder. The regression estimates of β, γ, δ in second-stage (**??**) 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,

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

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 (**??**). If the treatment Z_i is health insurance, outcome Y_i a measure of health outcomes, and the mediator D_i increased use of healthcare institutions, then this models the choice to visit the doctor more often in terms of health benefits relative to costs.[8] This makes the Roy model useful as a base case for CM, where selection-into-mediator may be driven by private information (unobserved by the researcher).

[8] If the choice is considers over a sum of outcomes, then a simple extension to a utility maximisation model maintains this same framework with expected costs and benefits. See **??**.

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 show Roy-selection in terms of unobserved and observed factors,

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

If selection follows a Roy model, and the mediator is ignorable, then unobserved benefits can play no part in selection. The only driver of selection are individuals' differences in costs (and not benefits). If there are any selection-into-D_i benefits unobserved to the researcher, then mediator ignorability cannot hold.

Proposition 1. *Suppose mediator selection follows a Roy model* (**??**)*, and selection is not fully explained by costs and observed gains. Then mediator 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 **??** 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 — whether the researcher observes them or not. The unobserved gains are unlikely to be fully captured by an observed control set $\boldsymbol{X}_i$, except in very special cases.

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. An unmotivated note saying "we conduct an informal mechanism analysis by controlling for this variable" or "we assume the mediator satisfies selection-on-observables" does not cut it here, and will lead to biased inference in practice.

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 (**??**) is not identified, because $U_{0,i}, U_{1,i}$ are unobserved, and lead to omitted variables bias.

$$\begin{aligned}\mathbb{E}\left[Y_i \mid Z_i, D_i, \boldsymbol{X}_i\right] &= \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\ &\quad + \underbrace{(1 - D_i)\,\mathbb{E}\left[U_{0,i} \mid D_i = 0, \boldsymbol{X}_i\right] + D_i \mathbb{E}\left[U_{1,i} \mid D_i = 1, \boldsymbol{X}_i\right]}_{\text{Unobserved confounding.}} \qquad (6)\end{aligned}$$

The CF approach models the contaminating terms in (**??**), avoiding the bias from omitting them in regression estimates. CF methods were first devised to correct for sample selection problems (**?**), and were extended to a general selection problem of the same form as **??** (**?**). 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 (IVs) 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 (**?**).

In the example of analysing health gains from health insurance (**?**), 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 Re-identification of Causal Mediation (CM) Effects

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

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

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

Assumption **??** is the monotonicity condition first used in an IV context (**?**). Here, it is assuming that people respond to treatment, Z_i, by consistently taking or refusing the mediator D_i (always or never-mediators), 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 **??** is that selection-into-mediator can be written as a selection model with ordered threshold crossing values that describe selection-into-D_i (**?**).

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

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 **?**,

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

where $U_i \coloneqq F_V\left(V_i \mid \boldsymbol{X}_i\right)$ follows a uniform distribution, 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 IV literature to the CM setting. The main conceptual difference is not assuming Z_i is a valid instrument for identifying the D_i on Y_i effect among compliers; it is using the selection model representation to correct for selection bias. See Appendix **??** for a validation of the general **?** equivalence result in a CM setting, with conditioning covariates $\boldsymbol{X}_i$.

Assumption CF–2. Selection on mediator benefits.

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

Assumption **??** 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 $U_{1,i}$). If there is confounding in Y_i, then it can be measured in D_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 $\boldsymbol{X}_i^{\mathrm{IV}}$ as one entry in the vector, and $\boldsymbol{X}_i^{-}$ as the remaining.

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

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

Assumption **??** is requiring at least one control variable guides selection-into-D_i — an 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* **??**, **??**, **??** *hold, then second-stage regression equation* (**??**) *is identified with a CF adjustment.*

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

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 separately identified in the first-stage* (**??**). *Proof: see Appendix* **??**.

Again, this set-up required no linearity assumptions, and treatment effects vary, because Z_i, D_i are categorical and $\beta, \gamma, \delta, \varphi(\boldsymbol{X}_i)$ vary with $\boldsymbol{X}_i$. 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 **?**, 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:

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

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, **?**). If we do not know what distribution class the errors follow, then λ_0, λ_1 can be estimated separately with semiparametric methods to avoid relying on parametric assumptions.

Theorem CF. If assumptions **??**, **??**, **??** hold, the ADE and AIE are identified as a function of the parameters in Proposition **??**.

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

where $\Gamma(p, p') = \mathbb{E}\left[F_V^{-1}(U_i \mid \boldsymbol{X}_i) - \mu_V \,\middle|\, p < U_i \leq p'\right] = \frac{p'\lambda_1(p') - p\lambda_1(p)}{p' - p}$ is the average unobserved net gains for those with unobserved costs between $p < p'$,[9] and $\overline{\pi} = \pi(1;\boldsymbol{X}_i) - \pi(0;\boldsymbol{X}_i)$ is the mediator complier score. Proof: see Appendix **??**.

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, and thus purge persistent bias in identifying CM effects.

This identification strategy is essentially a Marginal Treatment Effect approach (MTE, **?**) applied to a CM setting. Just as the semiparametric local IV approach uses variation in instruments to identify MTEs across the distribution of unobserved treatment take-up costs, this CF approach identifies CM effects across the distribution of unobserved mediator take-up costs. This connection to MTEs provides a conceptual bridge between the literature on IV structural causal effects and CM.

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

[9] The complier adjustment term was first written in this manner by **?** for an IV setting.

Figure 2: The CF Adjustment Addresses Persistent Bias in Conventional CM Estimates.

(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 normally distributed error terms (with both correlation and heteroscedasticity, further described in **??**). The black dashed line is the true value; orange is the distribution of conventional CM estimates from two-stage OLS (**?**), and blue estimates with a two-stage Heckman selection adjustment.

4 Control Function (CF) Estimation of CM Effects

A conventional approach to estimating CM effects involves a two-stage approach to estimating the ADE and the AIE: the first-stage (Z_i on D_i), and the second-stage (Z_i, D_i on Y_i). A CF approach is a simple and intuitive addition to this approach: including the CF terms λ_0, λ_1 in the second-stage regression to address selection-into-mediator.

This section presents two practical estimation strategies. First, I demonstrate how to estimate CM effects with an assumed distribution of error terms, focusing on the Heckman selection model as the leading case. Second, I consider a more flexible semi-parametric approach that avoids distributional assumptions — at the cost of semi-parametrically estimating the corresponding CFs. While both methods effectively address the selection bias issues detailed in previous sections, they differ in their implementation complexity, efficiency, and underlying assumptions.

4.1 Parametric CF

A parametric CF solves the identification problem by assuming a distribution for the unobserved error terms in the first-stage selection model, and modelling selection based on this distribution. The Heckman selection model is the most pertinent example, assuming the normal distribution for unobserved errors (**?**). A parametric CF using other distributions works in exactly the same manner, replacing the relevant density functions for an alternative distribution as needed. As such, this section focuses exclusively on the Heckman selection model.

The Heckman selection model assumes unobserved errors V_i follow a normal distribution, so estimates the first-stage using a probit model.

$$\Pr\left(D_i = 1 \,|\, Z_i, \boldsymbol{X}_i\right) = \Phi\big(\theta + \overline{\pi} Z_i + \boldsymbol{\zeta}' \boldsymbol{X}_i\big),$$

where $\Phi(.)$ is the cumulative density function for the standard normal distribution, and $\theta, \overline{\pi}, \boldsymbol{\zeta}$ are parameters estimated with maximum likelihood.

From this probit first-stage, we construct an estimate of the inverse Mills ratio terms to serve as the CFs. These terms capture the correlation between unobserved factors influencing both mediator selection and outcomes, when the errors are normally distributed.

$$\lambda_0(p') = -\frac{\phi(p')}{\Phi(p')}, \quad \lambda_1(p') = \frac{\phi(p')}{\Phi(p')}, \qquad \text{for } p' \in (0,1)$$

where $\phi(.)$ is the probability density function for the standard normal distribution.

Lastly, the second-stage is estimated with OLS, including the estimated CFs.

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

where $\widehat{\pi}\big(z'; \boldsymbol{X}_i\big) = \Phi\Big(\widehat{\theta} + \widehat{\overline{\pi}} z' + \widehat{\boldsymbol{\zeta}}' \boldsymbol{X}_i\Big)$ are the predictions from the probit first-stage.

The resulting ADE and AIE estimates are composed from sample estimates of the terms in Theorem **??**,

$$\widehat{\text{ADE}} = \widehat{\gamma} + \widehat{\delta}\, \overline{D}_i, \quad \widehat{\text{AIE}} = \widehat{\overline{\pi}} \left(\widehat{\beta} + \widehat{\delta}\, \overline{Z}_i + \big(\widehat{\rho}_1 - \widehat{\rho}_0\big) \Gamma\big(\widehat{\pi}(0; \boldsymbol{X}_i), \widehat{\pi}(1; \boldsymbol{X}_i)\big)\right)$$

where $\overline{D}_i = \frac{1}{N} \sum_{i=1}^N D_i$, $\overline{Z}_i = \frac{1}{N} \sum_{i=1}^N Z_i$, and $\Gamma(.,.)$ is the mean estimate of the complier adjustment term as a function of λ_1.

The standard errors for estimates can be computed using the delta method. Specifically, we account for both the sampling variability in the coefficient estimates and the fact that the

CFs themselves are estimated in the first-stage. This approach yields $\sqrt{n}$-consistent estimates when the underlying error terms follow a bivariate normal distribution — i.e., when λ_0, λ_1 and $\widehat{\pi}$ are correctly modelled by the probit first-stage. Errors can also be estimated by the bootstrap, including estimation of both the first and second-stage within each bootstrap iteration.

In practice, a parametric CF approach is simple to implement using standard statistical packages. The key advantage is computational simplicity and efficiency, particularly in moderate-sized samples. However, this comes at the cost of strong distributional assumptions. For example, if the error terms deviate substantially from joint normality, the estimates may be biased.[10]

4.2 Semi-parametric CF

For settings where researchers are not comfortable specifying a specific distribution for the error terms, a semi-parametric CF will nonetheless consistently estimate CM effects. This method maintains the same identification strategy but avoids assuming a specific parametric error distribution.

The semi-parametric approach begins with flexible estimation of the first-stage, non-parametrically estimating the mediator propensity score,

$$\pi\left(Z_i; \boldsymbol{X}_i\right) = \mathbb{E}\left[D_i \,|\, Z_i, \boldsymbol{X}_i\right],$$

where $\boldsymbol{X}_i$ must include the instrument(s) $\boldsymbol{X}_i^{\text{IV}}$. This can be estimated using flexible methods such as series approximation or kernel-based approaches, as long as the first-stage is estimated $\sqrt{n}$-consistently.[11]

Next the CFs, themselves, are estimated with semi-parametric methods. Consider the $D_i = 0$ subsample first.

$$\mathbb{E}\left[Y_i \,|\, Z_i, D_i = 0, \boldsymbol{X}_i\right] = \alpha + \gamma Z_i + \varphi\big(\boldsymbol{X}_i^-\big) + \rho_0 \lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big),$$

which gives a regression equation to estimate semi-parametrically, with the first-estimate estimate $\widehat{\pi}\left(Z_i; \boldsymbol{X}_i\right)$ plugged in. The linear parameters (α, γ and a linear approximation of nui-

[10] While this concern is immaterial in an IV setting estimating the LATE (**?**), it is pertinent in this setting as the CF extrapolates to a different group of compliers.

[11] If an estimate of the first-stage that is not $\sqrt{n}$-consistent is used (e.g., a modern machine learning estimator), then the resulting second-stage estimate will not be $\sqrt{n}$-consistent. This could be ameliorated by augmenting the approach with cross-fitting, and the appropriate Neyman orthogonal moments; **?** use this approach for one-sided selection problems, but (as fas as I am aware) there is no general double machine learning approach for CF methods for two-sided selection problems.

sance function $\varphi(.)$, $\boldsymbol{\varphi}$)[12] can be estimated with OLS, with λ_0 taking a flexible semi-parametric specification. An attractive option is a series estimator, such as a spline specification, as this estimates the function without assuming a functional form but maintains $\sqrt{n}$-consistency. Note that λ_0 can no longer be separated from ρ_0 in this semi-parametric approach; this is inconsequential because the complier adjustment term requires ρ_0, ρ_1 only be identified up to a constant.[13] Next, the $\rho_0\lambda_0$ function is extrapolated to the $D_i = 1$ side, identifying the remaining terms β, δ, and thus the ADE and AIE.

Return to the $D_i = 1$ subsample,

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

The same process extrapolates the series estimate of $\rho_1\lambda_1$ from the $D_i = 1$ sample to the $D_i = 0$ subsample, for another set of estimates for the ADE and AIE. Efficient estimates of CM effects then composes these two, with weights proportional to the variance of each side.

This approach achieves valid estimation of the CM effects, without specifying the distribution behind unobserved error terms, and achieves desirable properties as long as the first-stage correctly estimates the mediator propensity score, and the structural assumptions hold true. The standard errors for estimates can again be computed using the delta method, or estimated by the bootstrap — again, across both first and second-stages within each bootstrap iteration. Note that relying on propensity score estimation requires assumptions that can be found wanting in real-world settings; a common support condition for the mediator is required, and the nonparametric first-stage may become cumbersome if there are many control variables.

4.3 Simulation Evidence

The following simulation gives an example to show how these methods work 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 $i = 1, \ldots, N$, with $N = 1,000$ for this simulation.

$$Z_i \sim \text{Binom}\,(0.5)\,, \quad \boldsymbol{X}_i^- \sim N(4,1), \quad \boldsymbol{X}_i^{\text{IV}} \sim \text{Unif}\,(-1,1)\,, \quad (U_{0,i}, U_{1,i}, U_{C,i}) \sim N\,(\boldsymbol{0}, \boldsymbol{\Sigma})$$

$\boldsymbol{\Sigma}$ is the matrix of parameters which controls the level of confounding from unobserved costs and benefits.[14]

[12] Appropriate interactions between Z_i, D_i and $\boldsymbol{X}_i$ can also flexibly control for $\boldsymbol{X}_i$, again avoiding linearity assumptions.

[13] Appendix **??** explains these points in further detail.

[14] The correlation and relative standard deviations for $U_{0,i}, U_{1,i}$ affect how large selection bias in conventional CM estimates; correlation for these with unobserved costs $U_{C,i}$ does not particularly matter, though increased

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

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

Following **??**, these data have the following first and second-stage equations:

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

Treatment Z_i has a causal 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 Z_i, D_i have constant partial effects, the ATE depends on how many i choose to take the mediator and there is treatment effect heterogeneity.

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 conventional approach to CM does not give an estimate for how much of the ATE goes through mediator D, but is contaminated by selection bias thanks to the unobserved error terms.

I simulate this data generating process 10,000 times, using $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.75 & 0 \\ 0.75 & 2.25 & 0 \\ 0 & 0 & 0.25 \end{pmatrix}$,[15] and estimate CM effects with conventional CM methods (two-stage OLS) and the introduced CF methods. In this simulation $\Pr(D_i = 1) = 0.379$, and 65.77% of the sample are mediator compliers (for whom $D_i(0) = 0$ and $D_i(1) = 1$). This gives an ATE value of 2.60, ADE 1.38, and AIE 1.22, respectively.[16]

?? shows how these estimates perform, with a parametric CF approach, relative to the true value; **??** does the same for a semi-parametric approach. The OLS estimates' distribution do not overlap the true values for any standard level of significance; the distance between the OLS estimates and the true values are the underlying bias terms derived in **??**. The parametric CF approach perfectly reproduces the true values, as the probit first-stage correctly models the normally distributed error terms. The semi-parametric CF estimates correct conventional

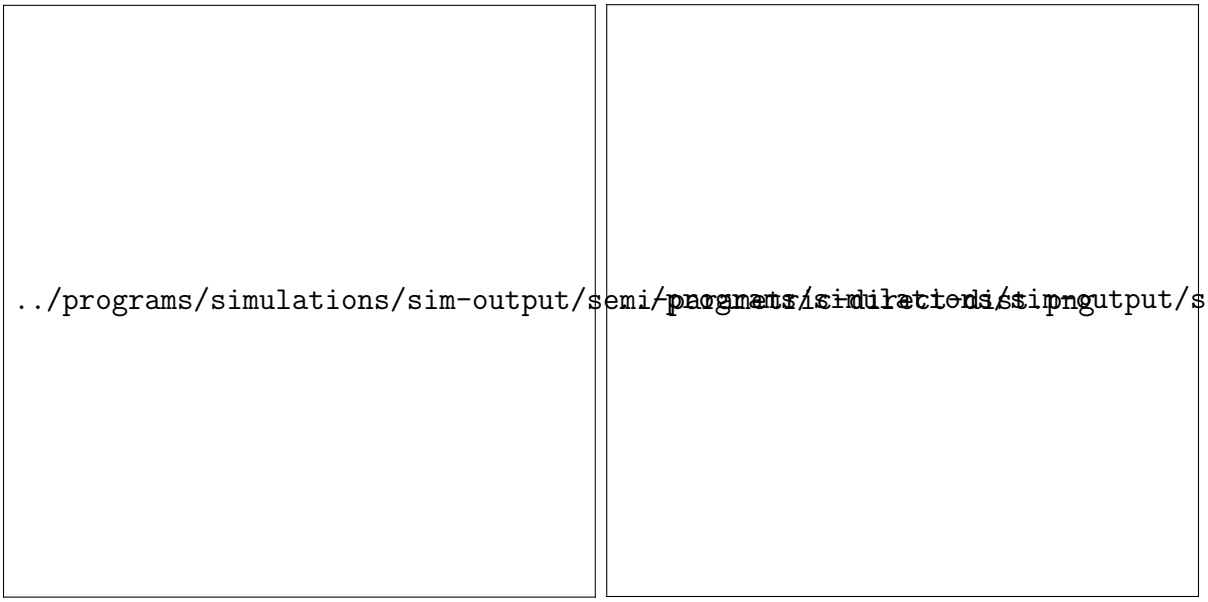
variance in unobserved costs makes estimates less precise for both OLS and CF methods.

[15] This choice of parameters has $\text{Var}(U_{0,i}) = 1, \text{Var}(U_{1,i}) = 2.25, \text{Corr}\left(U_{0,i}, U_{1,i}\right) = 0.5$ so that unobserved errors meaningfully confound conventional CM methods, with notable heteroscedasticity. Unobserved costs are uncorrelated with $U_{0,i}, U_{1,i}$ (although non-zero correlation would not meaningfully change the results), and $\text{Var}(U_{C,i}) = 0.25$ maintains uncertainty in unobserved costs.

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

Figure 3: Simulated Distribution of CM Effect Estimates, Semi-parametric versus OLS, 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. The black dashed line is the true value; orange is the distribution of conventional CM estimates from two-stage OLS (**?**), and green estimates with a two-stage semi-parametric CF.

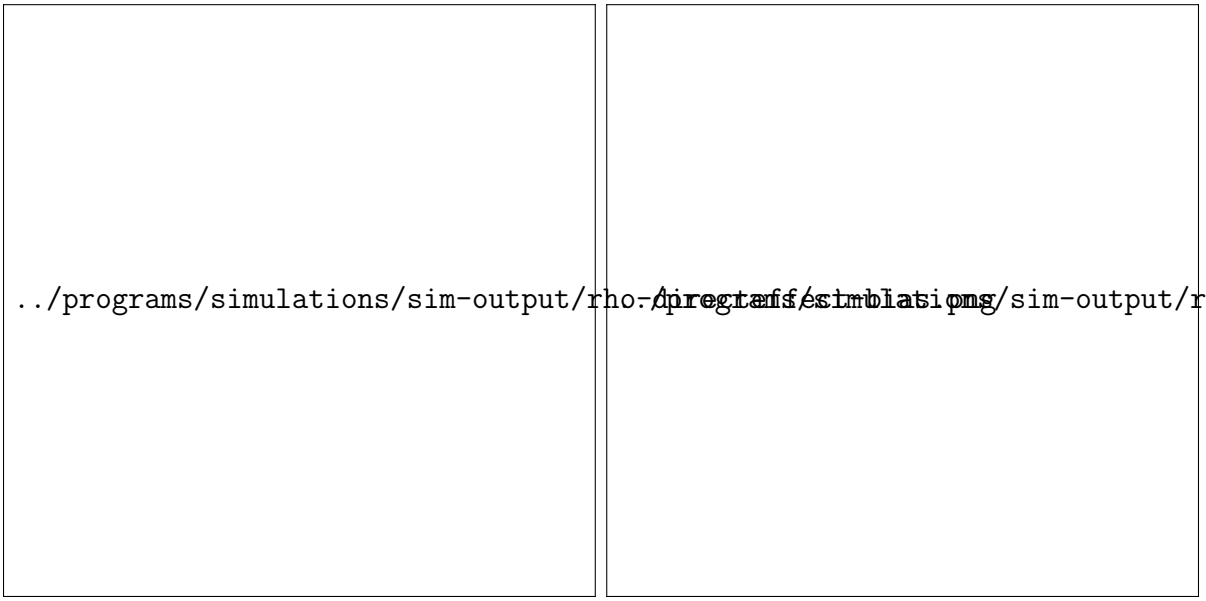
CM estimates, too, though exhibits some small-sample bias, which it to be expected with the involved non-parametric steps in realistic sample sizes.

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 $\text{Corr}\big(U_{0,i}, U_{1,i}\big) \in [-1, 1]$ and $\text{Var}\left(U_{0,i}\right) = 1, \text{Var}\left(U_{1,i}\right) \geq 0$. The true AIE values vary, because $D_i(Z_i)$ compliers have higher average values of $\big(U_{1,i} - U_{0,i}\big)$ as $\text{Corr}\big(U_{0,i}, U_{1,i}\big)$ increases. **??** 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 **??**). The CF approach improves on OLS estimates by correcting for bias, with confidence regions overlapping the true values.[17] This correction did not come for free: the standard errors are significantly greater in a CF approach than OLS. In this manner, this simulation shows the pros and cons of using the CF approach to estimating CM effects in practice.

[17]In the appendix, **??** shows the same simulation while varying $\text{Var}\left(U_{1,i}\right)$, with fixed $\text{Var}\left(U_{0,i}\right) = 1, \text{Corr}\big(U_{0,i}, U_{1,i}\big) = 0.5$. The conclusion is the same as for varying the correlation coefficient, ρ, in **??**.

Figure 4: CF Adjusted Estimates Work with Different Error Term Parameters.

(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, varying $\text{Corr}\left(U_{0,i}, U_{1,i}\right)$ values with $\text{Var}\left(U_{0,i}\right) = 1, \text{Var}\left(U_{1,i}\right) = 1.5$ fixed. The black dashed line is the true value, coloured points are points estimates for the respective data generated, and shaded regions are the 95% confidence intervals from 1,000 bootstraps each. Orange represents OLS estimates, blue the CF approach.

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 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 (**??**).

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 in natural experiment settings.

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 Previous Literature

Create a table in this section that surveys previous research which employs mediation methods while having a clear causal design for Z_i, but not D_i.

Paper	Field	Research Design for Z_i	Research Design for D_i	Selection bias?
Paper name 1.				

A.2 Identification in Causal Mediation

?, 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.

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

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, $\Pr(D_i(0) \neq D_i(1) \mid \boldsymbol{X}_i) = \mathbb{E}[D_i \mid Z_i = 1, \boldsymbol{X}_i] - \mathbb{E}[D_i \mid Z_i = 0, \boldsymbol{X}_i]$.

A.3 Bias in Causal Mediation (CM) Estimands

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

A.3.1 Bias in the Average Direct Effect (ADE)

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 $\boldsymbol{X}_i$ and $d' \in \{0, 1\}$.

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

And so,

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

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.

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

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

Collect everything together, as follows.

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

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

A.3.2 Bias in the Average Indirect Effect (AIE)

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 $\boldsymbol{X}_i$.

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

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

Let $z' \in \{0, 1\}$. Again, note the mean outcomes in terms of average potential outcomes,

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

So compose the selection bias term, as follows.

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

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

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.

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

Collect everything together, as follows.

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

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

A.4 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:

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

$U_{0,i}, U_{1,i}$ 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.

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

Since the ignorability assumption gives $\mathbb{E}\left[Z_i\eta_{z',i} \,|\, \boldsymbol{X}_i\right] = \mathbb{E}\left[Z_i \,|\, \boldsymbol{X}_i\right]\mathbb{E}\left[\eta_{z',i} \,|\, \boldsymbol{X}_i\right] = 0$, 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.

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

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:

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

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 (**?**). 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.

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

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.5 Roy Model and Sequential Ignorability

Proof of Proposition **??**.

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

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

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 $z', d' = 0, 1$

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

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.6 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 **?** *equivalence result, now including conditioning covariates* $\boldsymbol{X}_i$*, and is presented merely as a validation exercise.*

Assume condition monotonicity **??** 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 (**?**, 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 **?**. 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

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

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

A.7 Control Function (CF) Identification of the Second-stage

Proof of Proposition **??**. *This proof relies heavily on the notation and reasoning of* **?** *for an IV setting.*

By Assumption **??** (mediator monotonicity), selection-into-mediator can be represented as a threshold-crossing selection model.

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

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 **??** 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 crucial for the identification strategy, as it isolates the component of U_i that is related to selection-into-D_i.

The non-zero covariance condition in **??** 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 $\lambda_0\left(p'\right) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, p' < U_i\right]$. The control function $\lambda_0\left(p'\right)$ 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 $\lambda_1\left(p'\right) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, U_i \leq p'\right]$ for $p' \in (0,1)$, and $\rho_1 = \frac{\mathrm{Cov}\left(U_{1,i}, F_V^{-1}(U_i|\boldsymbol{X}_i)\right)}{\mathrm{Var}\left(F_V^{-1}(U_i|\boldsymbol{X}_i)\right)}$ is the corresponding projection coefficient.

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

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

This relationship ensures consistency in the CF approach across the $D_i = 0$ and $D_i = 1$ groups (**?**).

Assumption **??** (mediator take-up cost instrument $\boldsymbol{X}_i^{\mathrm{IV}}$) 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,

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

This gives the final result,

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

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 **?**), 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.8 Control Function (CF) Identification of the ADE and AIE

Proof of Theorem **??**.

Assume **??**, **??**, **??** hold. Then Proposition **??** 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.

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

This complier adjustment was first presented for an IV setting by **?**.

Collecting for the AIE,

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

A.9 Semiparametric Estimation of the ADE and AIE

Note the following, to express $\Gamma(p,p')$ in terms of either λ_0 or λ_1 for $p < p' \in (0,1)$.

$$
\begin{aligned}
\Gamma\left(p,p'\right) &= \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, p < U_i \leq p'\right] \\
&= \frac{p'\lambda_1\left(p'\right) - p\lambda_1\left(p\right)}{p' - p} \\
&= \frac{(1-p)\lambda_0\left(p\right) - (1-p')\lambda_0\left(p'\right)}{p' - p}
\end{aligned}
$$

The semiparametric approach cannot separate ρ_0 from λ_0, only estimating their composition $\rho_0\lambda_0(p')$ for $p' \in (0,1)$. Ultimately this does not matter, as it sufficient to identify ρ_0, ρ_1 up to a constant for the complier adjustment term, using either $\rho_0\lambda_0$ or $\rho_1\lambda_1$. For

$\pi_0 < \pi_1 \in (0,1)$,

$$\begin{aligned}(\rho_1 - \rho_0)\,\Gamma\big(\pi_0, \pi_1\big) &= (\rho_1 - \rho_0)\frac{\pi_1\lambda_1(\pi_1) - \pi_0\lambda_1(\pi_0)}{\pi_1 - \pi_0} \\ &= \left(1 - \frac{\rho_0}{\rho_1}\right)\frac{\pi_1\rho_1\lambda_1(\pi_1) - \pi_0\rho_1\lambda_1(p)}{\pi_1 - \pi_0} \\ &= \left(\frac{\rho_1}{\rho_0} - 1\right)\frac{(1-\pi_0)\rho_0\lambda_0(\pi_0) - (1-\pi_1)\rho_0\lambda_0(\pi_1)}{\pi_1 - \pi_0}.\end{aligned}$$

The fraction ρ_1/ρ_0 can be estimated with only the composition $\rho_0\lambda_0$ or $\rho_1\lambda_1$, thanks to the relation between λ_0, λ_1. For $p' \in (0,1)$,

$$\lambda_1(p') = \Longrightarrow \frac{\rho_0}{\rho_1} = \frac{\rho_0\lambda_0(p')}{\rho_1\lambda_1(p')}\left(\frac{1-p'}{p'}\right).$$

This can be estimated by estimating $\rho_1\lambda_1$ in the $D_i = 1$ subsample with splines, extrapolating the transformed $\rho_1\lambda_0(p') = -\rho_1\lambda_1(p')\left(\frac{p'}{1-p'}\right)$ to the $D_i = 0$ subsample, and calculating the regression coefficient for ρ_0/ρ_1 on $\rho_1\lambda_0(p')$.

Similarly, an approach starting with estimating $\rho_0\lambda_0$ in the $D_i = 0$ subsample gives a coefficient for ρ_1/ρ_0. An efficient estimate can compose these two, with weights proportional to the efficiency of each side.

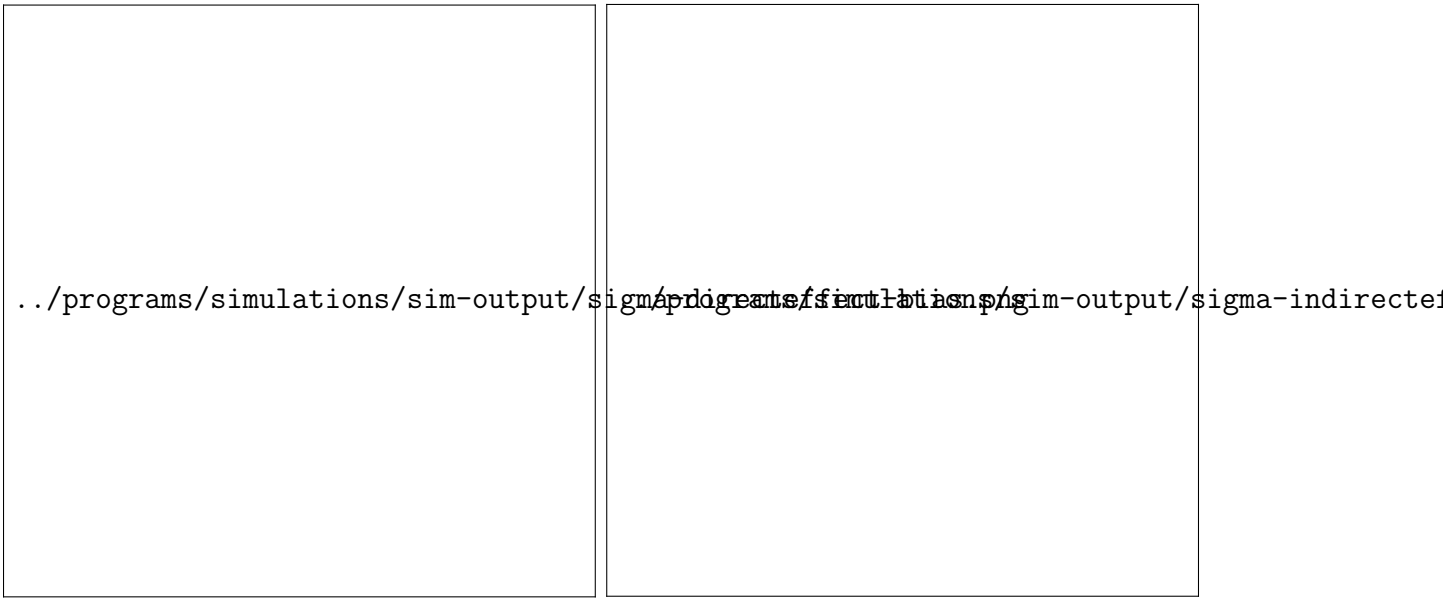
A.10 Control Function Simulation

A number of statistical packages, for the R language (**?**), made the simulation analysis for this paper possible.

- *Tidyverse* (**?**) collected tools for data analysis in the R language.
- *Splines* (**?**) allows semi-parametric estimation, using splines, in the R language.
- *Mediate* (**?**) automates the sequential-ignorability estimates of CM effects (**?**) in the R language.

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

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