Causal Mediation in Natural Experiments

Senan Hogan-Hennessy*
Economics Department, Cornell University†

First draft: 12 February 2025
This version: 25 April 2025

Work in Progress, newest version available here.

Abstract

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

Keywords: Direct/indirect effects, quasi-experiment, selection, control function.

JEL Codes: C21, C31.

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

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

Economists use natural experiments to credibly answer social questions, when an experiment was infeasible. Does access to health insurance lead to health improvements? Do transfer payments lead to measurable long-run economic gains? Quasi-experimental variation gives methods to answer these questions, but no indication of how these effects came about. Causal Mediation (CM) aims to estimate the mechanisms behind causal effects, by estimating how much of the treatment effect operates through a proposed mediator. For example, how much of the (causal) gain from a transfer payment came from individuals choosing to attend higher education? This paper shows that the conventional approach to estimating CM effects is inappropriate in a natural experiment setting, giving a theoretical framework for how bias operates, and 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 ignorable? Estimates for the average direct and indirect effects are contaminated by bias terms — selection bias plus group difference terms. I then show how this bias operates in an applied regression framework, with bias coming from a correlated error term. If individuals have been choosing to take or refuse a mediator based on expected costs and benefits (i.e., following a rational maximisation process) then sequential ignorability only holds if the set of observed control variable explains every conceivable source of mediator gains. 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 mediator ignorability is unlikely to hold true. This means investigating mechanisms by CM methods will lead to biased inference in natural experiment settings.

I consider an alternative approach to estimating CM effects, correcting for unobserved selection-into-mediator with a selection model. This approach solves the identification problem with structural assumptions for selection-into-mediator — mediator monotonicity and selection based on costs and benefits — and requires a valid cost instrument for mediator take-up. While these assumptions are strong, they are plausible in many applied settings.

Mediator monotonicity is in-line 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 in treatment to estimate causal effects. Access to a valid instrument is a string assumption, though is important to avoid further modelling assumptions; the most compelling example is using variation in mediator take-up cost as a first-stage instrument. This approach is not perfect in every setting: the structural assumptions are string, and are tailored to selection-into-mediator concerns pertinent to economic applications. This approach provides no harbour for estimating CM effects when a mediator is not ignorable, if these structural assumptions do not hold true.

The most popular approach to CM assumes that the original treatment, and the subsequent mediator, are both ignorable (Imai, Keele & Yamamoto 2010). This approach arose in the statistics literature, and is widely used in epidemiology, medicine, and psychology to estimate CM effects in observational studies. Assuming mediator ignorability (also known as selection-on-observables) conveniently ignores individuals' choice to take or refuse the mediator, by assuming they did so naïvely or the researcher observed everything that could have affected this decision. If a researcher is studying single-celled organisms in a laboratory, then it may make sense to study causal mechanisms with this approach; single-celled organisms would make simple decisions to take or refuse a treatment or mediator. On the other hand, social science researchers study social settings where humans make complex decisions based on costs, benefits, and preferences — all of which may not be observed fully by the researcher. Assuming a mediator is ignorable in such a setting would be naïve at best. In practice, the only setting where the mediator ignorability assumption is credible is using another natural experiment for the mediator. Given how hard it is to find random variation for one variable, it is a very limited setting to find another happening at the same time for the mediator of interest.

The applied econometrics literature has been hesitant to use explicit mediation analyses, but has picked up the practice of controlling for a mediator in an informal mechanism investigation (Blackwell, Ma & Opacic 2024). This practice is fundamentally a CM analysis, despite not being named so explicitly, so falls prey to the assumptions of explicit CM analyses. Indeed, a new strand of the econometric literature has developed estimators for CM effects under a variety of strategies to avoid this. This includes overlapping quasi-experimental research designs (Deuchert, Huber & Schelker 2019, Frölich & Huber 2017, Heckman & Pinto 2015), partial identification (Flores & Flores-Lagunes 2009), or a hypothesis test of full mediation through observed channels (Kwon & Roth 2024) — see Huber (2020) for an overview. The new literature has arisen in partial acknowledgement that a conventional selection-on-observables approach to CM in an applied setting can lead to biased inference, and needs alternative methods for credible inference. This paper makes this part explicit, showing exactly how a conventional approach to CM in a natural experiment can fail in practice, and warding the applied economics literature away from this approach to investigating mechanisms.

This paper considers the case when it is not credible to assume the mediator is ignorable, leveraging classic labour economic theory for selection-into-treatment to identify direct and indirect effects. This refers to settings where none of the natural experiment research designs in the previously cited papers apply (i.e., the mediator is not ignorable). A selection-on-observables approach to CM in this setting suffers from bias of the same flavour as classic selection bias (Heckman, Ichimura, Smith & Todd 1998), plus additional bias from group differences. The group differences-bias is a non-parametric version of bad controls bias, which has only previously been studied in a linear setting (Cinelli, Forney & Pearl 2024, Ding & Miratrix 2015).

Throughout, I use the Roy (1951) model as a benchmark for judging the Imai et al. (2010) mediator ignorability assumption in a natural experiment setting, and find it unlikely to hold in practice.[1] This motivates a solution to the identification problem inspired by classic labour

[1] An alternative method to estimate CM effects is ensuring treatment and mediator ignorability holds by a running randomised controlled trial for both treatment and mediator at the same time. This set-up has

economic work, which also uses the Roy model as a benchmark (Heckman 1979, Heckman & Honore 1990). I follow the lead of these papers by using a selection model approach to correct for the bias developed above. This approach assumes mediator monotonicity, exploiting the instrumental variables equivalence result in a mediation setting (Vytlacil 2002). Second, it assumes that mediator take-up is motivated by mediator costs and benefits so that first-stage errors inform second-stage unobserved confounding (Florens, Heckman, Meghir & Vytlacil 2008). Last, it requires a valid instrument for mediator take-up, to avoid relying on parametric assumptions on unobserved selection (Heckman & Navarro-Lozano 2004). This approach takes insights from the instrumental variables literature (Kline & Walters 2019), to account for selection and complier differences in a CM analysis. Doing so is related to using instruments to identify CM effects among instrument complier groups — as noted by Frölich & Huber (2017).[2] Using a selection model to estimate CM effects builds on the influential Imai et al. (2010) approach, marrying the CM literature with labour economic theory on selection-into-treatment for the first time.

This paper proceeds as follows. Section 1 introduces CM, and develops expressions for the bias in CM estimates in natural experiments. Section 2 describes this bias in applied settings with (1) a regression framework, (2) a setting with selection based on costs and benefits. Section 4 achieves identification by a control function approach, in the case that a mediator is monotone in the original treatment and a researcher observes exogenous variation in cost of mediator take-up, giving simulation evidence. Section 5 concludes.

1 Direct and Indirect Effects

Causal mediation decomposes causal effects into two channels, through a mediator (indirect effect) and through all other paths (direct effect). To develop notation for direct and indirect

been considered in the literature previously, in theory (Imai, Tingley & Yamamoto 2013, **?**) and in practice (Ludwig, Kling & Mullainathan 2011, Heckman, Pinto & Savelyev 2013).

[2]Indeed, this paper does not improve on selection model or control function methods, instead noting its applicability in this setting. See Frölich & Huber (2017) for the newest development of control function methods with instruments, and Imbens (2007) for a general overview of the approach.

effects, write Z_i for an exogenous binary treatment, D_i a binary mediator, and Y_i an outcome for individuals $i = 1, \ldots, n$.[3] The outcomes are a sum of their potential outcomes.[4]

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

Assume Z_i is ignorable.[5]

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

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

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

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

It common in the economics literature to assume that Z influences D in at most one direction, $\Pr\left(D_i(1) \geq D_i(0)\right) = 1$ — monotonicity (Imbens & Angrist 1994). I assume monotonicity (and its conditional variant) holds through-out to simplify notation.[6]

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

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

Z affects outcome Y directly, and indirectly via the $D(Z)$ channel, with no reverse

[3] Other literatures use different notation. For example, Imai et al. (2010) write T_i, M_i, Y_i for the randomised treatment, mediator, and outcome, respectively. I use the Z_i, D_i, Y_i instrumental variables notation, more familiar in empirical economics (Angrist, Imbens & Rubin 1996).

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

[5] 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.

[6] Assuming monotonicity also brings closer to the IV notation, and has other beneficial implications in this setting (see Section 4).

causality. Figure 1 visualises the design, where the direction arrows denote the causal direction. CM aims to decompose the ATE of $Z \to Y$ into these two separate pathways:

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

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

Estimating the AIE answers the following question: how much of the causal effect $Z \to Y$ goes through the D channel? If a researcher is studying the health gains of access to 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 (Finkelstein, Taubman, Wright, Bernstein, Gruber, Newhouse, Allen, Baicker & Group 2012). Estimating the ADE answers the following equation: how much is left over after accounting for the D channel?[7] For the health insurance example, how much of the health insurance effect is a direct effect, other than increased healthcare usage — e.g., long-term effects of lower medical debt, or less worry over health shocks. An instrumental variables approach assumes this direct effect is zero for everyone — that health insurance increased health only via healthcare usage (i.e., the exclusion restriction). CM is a similar, yet distinct, framework attempting to explicitly model the direct effect, and not assuming it is zero.

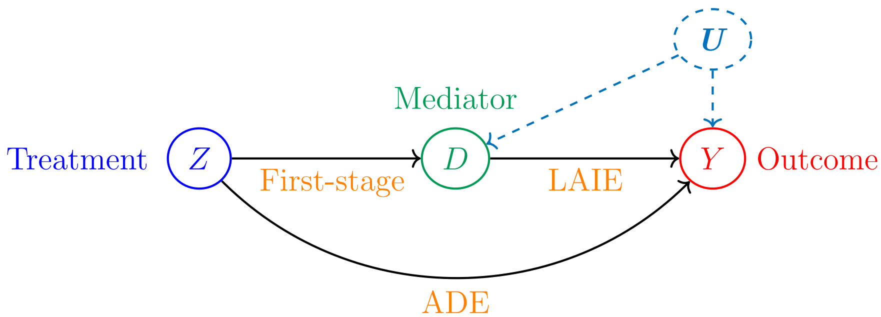
The ADE and AIE are not separately identified without further assumptions.

1.1 Identifying Causal Mediation (CM) Effects

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

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

Figure 1: Structural Causal Model for Causal Mediation.



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

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

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

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

Sequential ignorability assumes that the initial treatment Z_i is ignorable conditional on $\boldsymbol{X}_i$. It then also assumes that, after Z_i is assigned, that D_i is ignorable conditional on $\boldsymbol{X}, Z$. In addition, a common support condition for both Z_i, D_i (across $\boldsymbol{X}_i$) is necessary. If sequential ignorability, 1(1) and 1(2), holds then the ADE and AIE are identified by two-stage mean differences, after conditioning on $\boldsymbol{X}_i$.[8]

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

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

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

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 (Imai et al. 2010). While this is the most common approach in the applied literature, I do not assume the linear model. Linearity assumptions are unnecessary to my analysis; it suffices to note that heterogeneous treatment effects and non-linear confounding would bias OLS estimates of CM estimands in the same manner that is well documented elsewhere (see e.g., Angrist 1998, Słoczyński 2022). This section focuses on problems that plague CM by selection-on-observables, regardless of estimation method.

1.2 Bias in Causal Mediation (CM) Estimates

Applied research may use a natural experiment to study settings where treatment Z_i is ignorable, justifying assumption 1(1). Rarely does research relying on a quasi-experimental research design employ an additional, overlapping identification design for D_i to justify assumption 1(2) as part of the analysis. One might consider conventional CM methods in such a setting to learn about the mechanisms behind the causal effect $Z \to Y$ under study. This approach leads to biased estimates, and contaminates inference regarding direct and indirect effects.

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

Proof. See Subsection A.2 for the proof. Below I present the relevant selection bias and group difference terms, omitting the conditional on $\boldsymbol{X}_i$ notation for brevity. □

For the direct effect: CM estimand = ADE + selection bias + group differences.

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

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

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

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 $\boldsymbol{X}_i$.[9]

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

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

These selection bias terms would equal zero if the mediator had been ignorable 1(2), but do not necessarily average to zero if not.[10]

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

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

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

This group differences term in the AIE arises because the selection-on-observables approach does not account for complier differences.[11]

2 Causal Mediation (CM) in Applied Settings

In this section, I further develop the issue of selection in CM estimates. First, I show the non-parametric bias terms from above can be written as omitted variables bias in a regression framework. Second, I show how selection bias operates in an applied model for selection into a mediator based on costs and benefits.

[10] The selection-on-observables approach could, instead, focus on the average effect on treated populations — as do Keele, Tingley & Yamamoto (2015). This runs into a problem of comparisons: CM estimates would give average effects on different treated groups. The CM estimand for the ADE on treated gives the ADE local to the $Z_i = 1$ treated group, and for the ADE local to compliers with $D_i = 1$. In this way, these ADE and AIE on treated terms are not comparable to each other, so I focus on the true averages to avoid these misaligned comparisons.

[11] If the mediator had been ignorable, the complier indirect effect would equal the AIE; this differences term represents the difference between the two when it is not.

2.1 Regression Framework

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

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

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

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

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

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

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

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

First-stage (3) is identified, with $\phi + \zeta(\boldsymbol{X}_i)$ the intercept, and $\overline{\pi}$ the first-stage compliance rate (which may depend on $\boldsymbol{X}_i$). Second-stage (4) has the following definitions, and is not identified thanks to omitted variables bias.[12]

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

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

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

[12] See Subsection A.3 for the derivation.

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

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

The ADE and AIE are averages of these regression coefficients.

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

The ADE is a simple sum of the coefficients, while the AIE includes a group differences term because it only refers to the mediator compliers.[13]

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 (4) give unbiased estimates only if D_i is also conditionally ignorable: $D_i \perp\!\!\!\perp \boldsymbol{U}_i$. If not, then estimates of CM effects suffer from omitted variables bias from failing to adjust for the unobserved confounder, $\boldsymbol{U}_i$.

2.2 Selection on Costs and Benefits

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

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

[13] Section 3 returns to the issue of accounting for mediator compliers in identifying the average indirect effect.

model has i taking the mediator if the benefits exceed the costs,

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

The Roy model provides an intuitive framework for analysing selection mechanisms because it captures the fundamental economic principle of decision-making based on costs and benefits in terms of the outcome under study (Roy 1951, Heckman & Honore 1990). If the outcome Y_i is a measure of income, and the mediator a choice of taking education, then it models an individual choice to attend more education in terms of gaining a higher income compared to the costs.[14] This makes it particularly useful as a base case for CM, where selection into the mediator may be driven by private information (unobserved by the researcher). By using the Roy model as a benchmark, I explore the practical limits of the mediator ignorability assumption.

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

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

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

Definition 2. *Suppose mediator selection follows a Roy model (5), and selection is not fully explained by costs and observed gains. Then sequential ignorability does not hold.*

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

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

See Subsection A.4 for the proof.

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

3 Solving Identification with a Selection Model

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 example would yield unbiased estimates for the ADE and AIE. A selection model 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 reqression equation (4) is not identified, because $U_{0,i}$ and $U_{1,i}$ are unobserved.

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

The following assumptions are sufficient to model the correlated error terms, identifying β, γ, δ, and thus both the ADE and AIE.

Needs more story telling of how this happens.

Writing plan: set up a sequence of assumption X $\implies$ implcation X. Then a paragraph underneath, explaining what it is doing. See Kline & Walters (2019) for an excellent example of this.

Additionally, 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 rows.

Definition 3. *Control function assumptions.*

$$\Pr\left(D_i(1) \geq D_i(0) \mid \boldsymbol{X}_i\right) = 1 \tag{6}$$

$$D_i \perp\!\!\!\perp Y_i(.,.) \;\Big|\; \boldsymbol{X}_i^{-}, K_i \tag{7}$$

$$\boldsymbol{X}_i^{IV} \text{ satisfies } \frac{\partial}{\partial \boldsymbol{X}_i^{IV}} \left[\mu_1(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i)\right] = 0 < \frac{\partial}{\partial \boldsymbol{X}_i^{IV}} \mathbb{E}\left[D_i(z') \mid \boldsymbol{X}_i\right], \text{ for } z' = 0, 1. \tag{8}$$

Assumption 4(9) is the (conditional) monotonicity assumption (Imbens & Angrist 1994), which is untestable but acceptable in many empirical applications. Assumption 4(10) is the control function assumption, assuming that first-stage unobserved heterogeneity explains second-stage selection into D_i. Assumption 4(11) is assuming that 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 first-stage). The exclusion restriction is untestable, and must be guided by domain-specific knowledge; strength of the first-stage is testable, and must be justified with data by methods common in the instrumental variables literature.

Writing here about how a Heckman (1974) selection model purges selection bias.

Control function assumption + instrument.

Lemma: Under assumptions CF(1, 2, 3), mean POs are identified in the way written in Kline Walters (2019). Needs an appendix proof.

This is exploiting ideas from selection models + marginal TEs to identify this system, including using the selection model to identify the mediator compilers' effect. Indeed, mediation estimates already do a two-step procedure; it is a minor adjustment to include a CF in the second-stage, to guard against selection-on-gains (chief among which is the Roy model).

Monotonicity gives selection model representation, $D = 1\phi(Z, X) > V$ which can be transformed into $D = 1\pi(Z, X) > U$ for $U_i = F_V(V_i)$ $unif(0, 1)$. CF assumption connects first-stage and second-stage errors, by assuming that $Cov(U_i, U_1), Cov(U_i, U_0) > 0$. Instrument separately identifies the propensity score (not technically required but needed for efficiency).

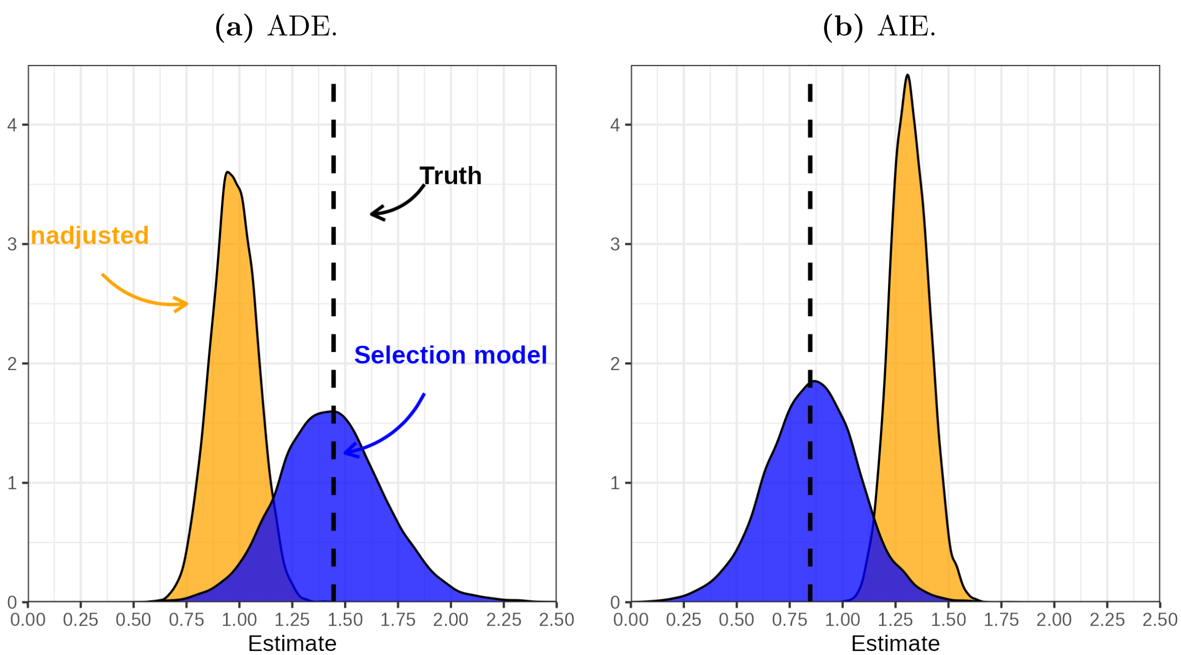
Under these assumptions, outcome regression has the following form.

$$E[Y|Z, D, X] = \ldots + \rho_0 \lambda_0(\pi(Z_i, X_i)) + \rho_1 \lambda_1(\pi(Z_i, X_i))$$

Where λ_0, λ_1 are the control functions. For $J(u) = F^-1_V(u) - E[F^-1_V(U_i)]$

This is a special case of Heckman (1980), notation from Kline Walters (2019).

Figure 2: Simulated Distribution of CM Effect Estimates.



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

4 Selection-Adjusted Estimation of CM Effects

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

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

4.1 Old writing

Write K_i for the expected values in predicting the mediator with observed data $Z_i, \boldsymbol{X}_i$.

$$K_i = D_i - \mathbb{E}\left[D_i \,|\, Z_i, \boldsymbol{X}_i\right]$$

Additionally, 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 rows.

Definition 4. *Control function assumptions.*

$$\Pr\left(D_i(1) \geq D_i(0) \,|\, \boldsymbol{X}_i\right) = 1 \tag{9}$$

$$D_i \perp\!\!\!\perp Y_i(.,.) \,\Big|\, \boldsymbol{X}_i^-, K_i \tag{10}$$

$$\boldsymbol{X}_i^{IV} \text{ satisfies } \frac{\partial}{\partial \boldsymbol{X}_i^{IV}}\left[\mu_1(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i)\right] = 0 < \frac{\partial}{\partial \boldsymbol{X}_i^{IV}}\mathbb{E}\left[D_i(z') \,|\, \boldsymbol{X}_i\right], \text{ for } z' = 0, 1. \tag{11}$$

Assumption 4(9) is the (conditional) monotonicity assumption (Imbens & Angrist 1994), which is untestable but acceptable in many empirical applications. Assumption 4(10) is the control function assumption, assuming that first-stage unobserved heterogeneity explains second-stage selection into D_i. Assumption 4(11) is assuming that 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 first-stage). The exclusion restriction is untestable, and must be guided by domain-specific knowledge; strength of the first-stage is testable, and must be justified with data by methods common in the instrumental variables literature.

K_i serves as a control function in this setting.

Theorem 2. *If the control function assumptions hold, then the mean potential differences (and thus CM effects) are identified by a control function approach. For each $z', d' = 0, 1$,*

$$\mathbb{E}\left[Y_i \,\middle|\, Z_i = 1, D_i = d', \boldsymbol{X}_i^-, K_i\right] - \mathbb{E}\left[Y_i \,\middle|\, Z_i = 0, D_i = d', \boldsymbol{X}_i^-, K_i\right] = \mathbb{E}\left[Y_i(1, d') - Y_i(0, d') \,\middle|\, \boldsymbol{X}_i^-\right]$$

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

Proof. Special case of Florens et al. (2008, Theorem 1), Imbens & Newey (2009, Theorem 3). □

Assumption 4(9) guarantees that mediator can be represented by a selection model (Vytlacil 2002), $D_i(.) = \mathbb{1}\{\overline{\mu}(., ; \boldsymbol{X}_i) \geq K_i\}$, for some function $\overline{\mu}$. Assumption 4(10) connects the first and second-stages for identification. Assumption 4(11) separately identifies the control function to identify the second-stage. This approach exploits the fact that the bias terms, coming from correlated errors in Subsection 2.1, can be modelled by the first-stage regression and included as controls in the second-stage.

If the underlying selection model had been a Roy model, the control function approach captures the unobserved benefits to taking mediator (independent of observed controls), and thus driving take-up of the mediator. In this case, $K_i = U_{C,i} - (U_{1,i} - U_{0,i})$, so the independence conditions follows. By incorporating the control function from the first-stage model, the approach adjusts for the unobserved confounding from unobserved gains, $U_{1,i} - U_{0,i}$. By contrast, assuming the mediator was ignorable would have been assuming that there are no unobserved benefits to the mediator take-up, so that there is no bias in the second-stage to account for.

The instrument is key to avoid distributional assumptions on the unobserved errors terms. In the Roy model, the exclusion restriction can be satisfied in one key way: having an instrument for cost of mediator take up μ_C. If the instrument $\boldsymbol{X}_i^{\text{IV}}$ enters the cost function μ_C, and not the benefits function $\mu_1 - \mu_0$, then it satisfies the exclusion restriction. In an applied world, $\boldsymbol{X}_i^{\text{IV}}$ can be data that explain cost differences in taking D_i, unrelated to other demographic information. If a researcher is looking into higher education as a proposed mediator, then data which explains different costs of attending university (unrelated to education gains) can serve this role. This is the logic behind the Card (1993) distance-instrument, and can be extended to a CM setting with education as the mediator.

4.2 Estimation

In practice, the approach relies on estimating the control function K_i, then including this in the second-stage as a control, and accounting for the estimation error for these in the standard errors. These reliances come with major concerns. First, it is imperative that the control function is estimated correctly, so it is necessary to employ a non-parametric approach to estimate the first-stage. Second, the error terms enters the second-stage (4) linearly, but is an unknown function (possibly non-linear) of the control function; thus, the second-stage must be estimated semi-parametrically.[15] Lastly, the standard errors must account for estimation uncertainty in the above two non-parametric steps.

These concerns are worth noting, because non-parametric regression is computationally demanding, and requires large samples for estimator convergence. Furthermore, these are estimated in two steps, so that the concerns are of greater importance. Otherwise, small sample bias properties could even dominate the bias terms identified in Theorem 1.[16] It is beyond the scope of this paper to develop the optimal procedure here, but these concerns are important. For applied research aiming to estimate CM effects, the control function method

[15] In practice this can be done by adding a polynomial for the estimated control function into the outcome regression, or a splines approach, etc.

[16] See (Imbens & Newey 2009, Section 6) for a full discussion of the asymptotic theory of a control function estimator.

is only appropriate in extremely large sample sizes, such as applications using administrative sources or biobanks.

With these concerns in mind, I propose the following method to estimate CM effects with a control function approach:

1. Estimate the first-stage, $\mathbb{E}\left[D_i \,\middle|\, Z_i, \boldsymbol{X}_i^{\text{IV}}, \boldsymbol{X}_i^{-}\right]$ with a non-parametric estimator (e.g., a probability forest, or fully interacted OLS specification).

2. Calculate estimates of the control function:

$$\widehat{K}_i = D_i - \widehat{\mathbb{E}}\left[D_i \middle| Z_i, \boldsymbol{X}_i^{\text{IV}}, \boldsymbol{X}_i^{-}\right].$$

3. Estimate the second-stage with OLS (including an interaction term between Z_i and D_i), and a semi-parametric regressor of the control function.

$$\mathbb{E}\left[Y_i \middle| Z_i, D_i, \boldsymbol{X}_i^{-}, \widehat{K}_{0,i}, \widehat{K}_i\right] = \beta D_i + \gamma Z_i + \delta Z_i D_i + l\left(\widehat{K}_i\right)$$

   $l(.)$ is a nuisance function with unknown form, so can be approximated with a semi-parametric spline specification, for example.

4. Calculate the ADE and AIE estimates from the first and second-stages.

$$\begin{aligned}
\widehat{\text{ADE}} = \;& \mathbb{E}\left[\widehat{\mathbb{E}}\left[Y_i \middle| Z_i = 1, D_i, \boldsymbol{X}_i^{-}, \widehat{K}_i\right] - \widehat{\mathbb{E}}\left[Y_i \middle| Z_i = 0, D_i, \boldsymbol{X}_i^{-}, \widehat{K}_i\right]\right] \\
\widehat{\text{AIE}} = \;& \mathbb{E}\left[\begin{array}{l} \left(\widehat{\mathbb{E}}\left[D_i \middle| Z_i = 1, \boldsymbol{X}_i^{\text{IV}}, \boldsymbol{X}_i^{-}\right] - \widehat{\mathbb{E}}\left[D_i \middle| Z_i = 1, \boldsymbol{X}_i^{\text{IV}}, \boldsymbol{X}_i^{-}\right]\right) \\ \times \left(\widehat{\mathbb{E}}\left[Y_i \middle| Z_i, D_i = 1, \boldsymbol{X}_i^{-}, \widehat{K}_i\right] - \widehat{\mathbb{E}}\left[Y_i \middle| Z_i, D_i = 0, \boldsymbol{X}_i^{-}, \widehat{K}_i\right]\right) \end{array}\right]
\end{aligned}$$

5. Bootstrap across the previous steps, to calculate standard errors for the respective ADE and AIE estimates.

4.3 Simulation Evidence

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

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

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

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

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

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

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

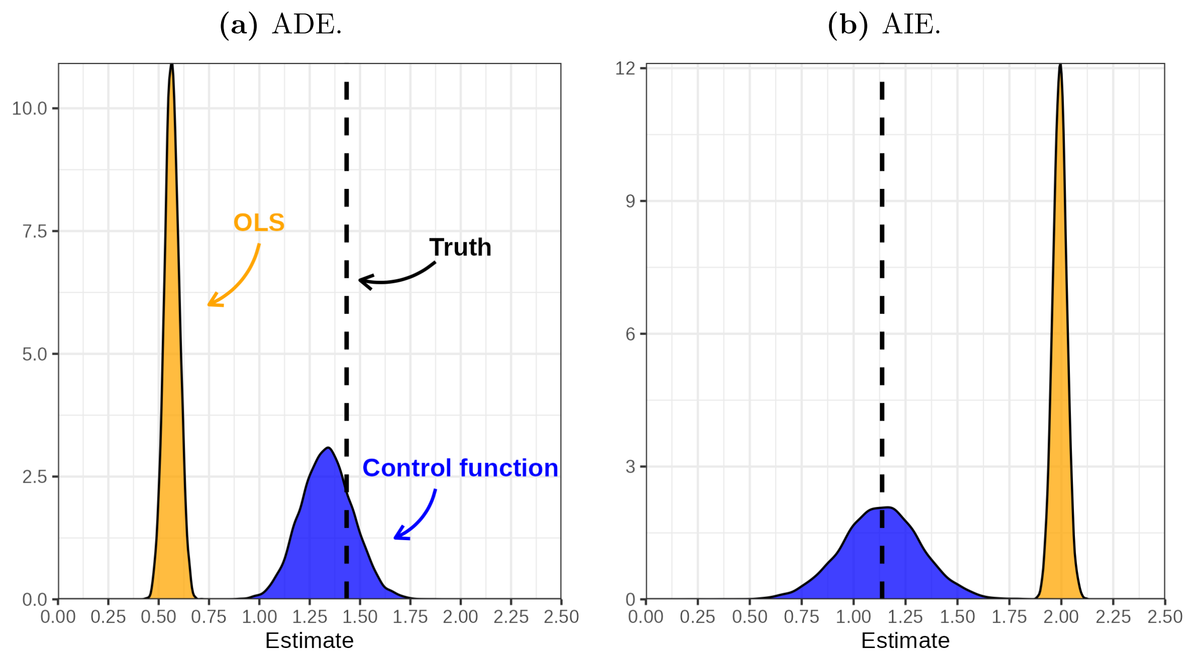
Z_i has an effect on outcome Y_i, and it operates partially through mediator D_i. Outcome mean $\mu_{D_i}(Z_i; .)$ contains an interaction term, $Z_i D_i$, so while both Z_i and D_i have constant partial effects, the ATE depends on how many i choose to take the mediator. In this simulation $\Pr(D_i = 1) = 0.437$, and 65.29% of the sample are mediator compliers (where $D_i(1) = 1$ and $D_i(0) = 0$). This gives an ATE ($Z \to Y$) value of 2.58, ADE 1.44, and AIE 1.13, respectively.[17]

After Z_i is assigned, i chooses to take mediator D_i by considering the costs and benefits — which vary based on Z_i, demographic controls $\boldsymbol{X}_i$, and the (non-degenerate) unobserved error terms $U_{i,0}, U_{1,i}$. As a result, sequential ignorability does not hold; the mediator is not conditionally ignorable. Thus, a standard OLS (selection-on-observables) approach to

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

CM does not give an estimate for how much of the $Z \to Y$ ATE goes through mediator D. Instead, the OLS approach gives biased inference.

Figure 3: Simulated Distribution of CM Effect Estimates.



Note: These figures show the empirical density of point estimates, for 10,000 replications of the data generating process described above. The black dashed line is the true value; orange is the distribution of naïve OLS estimates, and blue the control function approach.

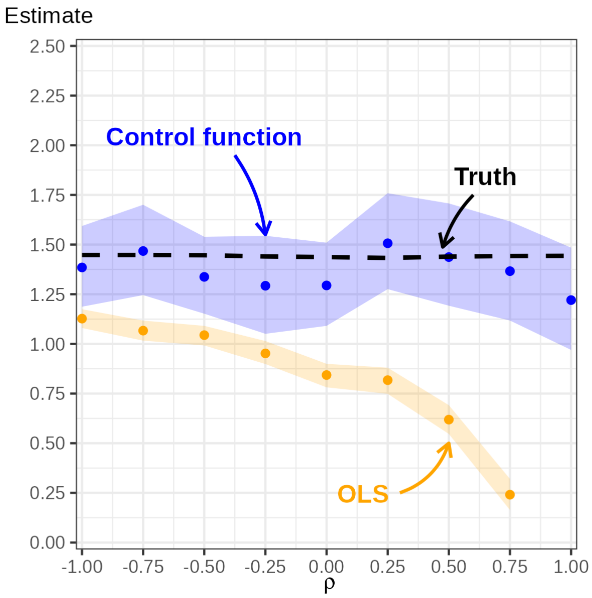
The bias in OLS estimates comes from the unobserved error terms being related. Figure 3 shows the distribution of bootstrapped point estimates in this simulation, showing OLS against the control function approach. The OLS approach implicitly assumes that the mediator is ignorable (when it is not), so its point estimates under and over-estimate the true ADE and AIE, respectively. The distance between the OLS estimates and the true values are the underlying bias terms derived in Theorem 1. In this data generating process, the OLS confidence interval do not overlap the true values for any standard level of significance. The control function approach exhibits bias, though the 95% confidence intervals cover the truth.

The error terms determine the bias in OLS estimates of the ADE and AIE, so the bias varies for different values of the error-term parameters $\rho \in [-1, 1]$ and $\sigma_0, \sigma_1 \geq 0$.[18] Figure 4

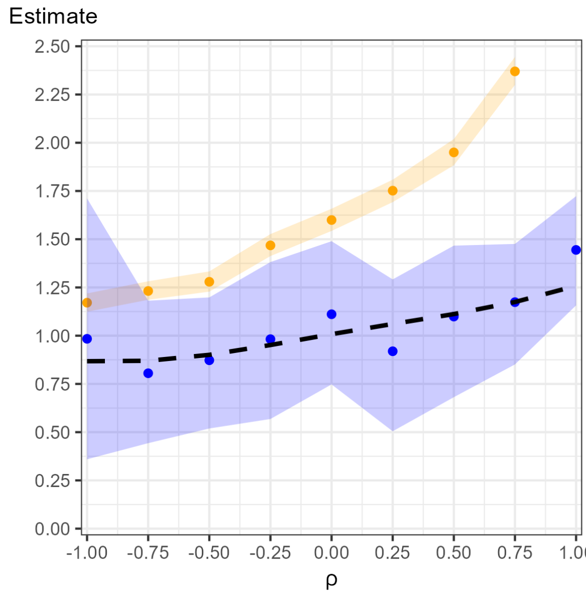
[18]Indeed, this setting has error terms following a bivariate normal distribution, so the canonical Heckman (1974) selection model would produce the most efficient estimates by maximum likelihood. The control

Figure 4: Point Estimates of CM Effects, OLS versus Control Function, varying ρ values with $\sigma_0 = 1, \sigma_1 = 2$ fixed.

(a) ADE.



(b) AIE.



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

shows control function estimates against estimates calculated by standard OLS, showing 95% confidence intervals calculated from 1,000 bootstraps. The point estimates of the control function do not exactly equal the true values, as they are estimates from one simulation (not averages across many simulations, as in Figure 3). The control function approach improves on OLS estimates by correcting for bias, with confidence regions overlapping the true values.[19,20] This correction did not come for free: the standard errors are significantly greater in a control function approach than OLS. Standard errors on the AIE are larger than those for the ADE, because the AIE estimates are first-stage times second-stage estimates, so standard

function approach avoids this assumption, and bias from breaking it, by relying on an instrument.

[19]The code behind this simulation estimates the first-stage with an interacted OLS specification, and splines included for the continuous regressor $\boldsymbol{X}_i^-$. The second-stage is an OLS specification, including the control function with a spline specification.

[20]In the appendix, Figure A1 shows the same simulation while varying σ_1, with fixed $\sigma_0 = 1, \rho = 0.5$. The conclusion is the same as for varying the correlation coefficient, ρ, in Figure 4.

errors account for uncertainty in both estimates multiplied. In this manner, this simulation shows the pros and cons of using the control function approach to estimating CM effects in practice.

5 Summary and Concluding Remarks

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

This paper has contributed to the growing CM literature in economics, integrating labour economic theory for select-into-treatment as a way of judging 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 control function as an additional robustness in the growing double robustness literature (Farbmacher, Huber, Lafférs, Langen & Spindler 2022, Bia, Huber & Lafférs 2024).

This paper has not lit the way for applied researchers to use CM methods unabashedly, with or without a control function adjustment. The structural assumptions are strong and large sample sizes are needed; if the assumptions are broken, then the control function method does not improve on a naïve selection-on-observables approach to CM estimates. 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 exogenous variation in mediator take-up costs. In these cases, this paper opens the door to identifying mechanisms behind treatment effects in natural experiment settings.

References

Angrist, J. D. (1998), 'Estimating the labor market impact of voluntary military service using social security data on military applicants', *Econometrica* **66**(2), 249–288. 8

Angrist, J. D. (2022), 'Empirical strategies in economics: Illuminating the path from cause to effect', *Econometrica* **90**(6), 2509–2539. 14

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American statistical Association* **91**(434), 444–455. 5

Angrist, J. D. & Pischke, J.-S. (2009), *Mostly harmless econometrics: An empiricist's companion*, Princeton university press. 9, 14

Bia, M., Huber, M. & Lafférs, L. (2024), 'Double machine learning for sample selection models', *Journal of Business & Economic Statistics* **42**(3), 958–969. 24

Blackwell, M., Ma, R. & Opacic, A. (2024), 'Assumption smuggling in intermediate outcome tests of causal mechanisms', *arXiv preprint arXiv:2407.07072* . 3

Card, D. (1993), 'Using geographic variation in college proximity to estimate the return to schooling'. `https://doi.org/10.3386/w4483`. 19

Cinelli, C., Forney, A. & Pearl, J. (2024), 'A crash course in good and bad controls', *Sociological Methods & Research* **53**(3), 1071–1104. 3, 10

Deuchert, E., Huber, M. & Schelker, M. (2019), 'Direct and indirect effects based on difference-in-differences with an application to political preferences following the vietnam draft lottery', *Journal of Business & Economic Statistics* **37**(4), 710–720. 3

Ding, P. & Miratrix, L. W. (2015), 'To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias', *Journal of Causal Inference* **3**(1), 41–57. 3, 10

Eisenhauer, P., Heckman, J. J. & Vytlacil, E. (2015), 'The generalized roy model and the cost-benefit analysis of social programs', *Journal of Political Economy* **123**(2), 413–443. 13

Farbmacher, H., Huber, M., Lafférs, L., Langen, H. & Spindler, M. (2022), 'Causal mediation analysis with double machine learning', *The Econometrics Journal* **25**(2), 277–300. 24

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. & Group, O. H. S. (2012), 'The oregon health insurance experiment: Evidence from the first year*', *The Quarterly Journal of Economics* **127**(3), 1057–1106. **URL:** *https://doi.org/10.1093/qje/qjs020* 6

Florens, J.-P., Heckman, J. J., Meghir, C. & Vytlacil, E. (2008), 'Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects', *Econometrica* **76**(5), 1191–1206. 4, 18

Flores, C. A. & Flores-Lagunes, A. (2009), 'Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness'. 3

Frölich, M. & Huber, M. (2017), 'Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **79**(5), 1645–1666. 3, 4

Heckman, J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica: journal of the econometric society* pp. 679–694. 15, 22

Heckman, J., Ichimura, H., Smith, J. & Todd, P. (1998), 'Characterizing selection bias using experimental data', *Econometrica* **66**(5), 1017–1098. 3, 9, 10

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica: Journal of the econometric society* pp. 153–161. 4

Heckman, J. J. & Honore, B. E. (1990), 'The empirical content of the roy model', *Econometrica: Journal of the Econometric Society* pp. 1121–1149. 4, 13

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

Heckman, J. J. & Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation 1', *Econometrica* **73**(3), 669–738. 11

Heckman, J. & Navarro-Lozano, S. (2004), 'Using matching, instrumental variables, and control functions to estimate economic choice models', *Review of Economics and statistics* **86**(1), 30–57. 4

Heckman, J., Pinto, R. & Savelyev, P. (2013), 'Understanding the mechanisms through which an influential early childhood program boosted adult outcomes', *American economic review* **103**(6), 2052–2086. 4

Huber, M. (2020), 'Mediation analysis', *Handbook of labor, human resources and population economics* pp. 1–38. 3

Huber, M., Hsu, Y.-C., Lee, Y.-Y. & Lettry, L. (2020), 'Direct and indirect effects of continuous treatments based on generalized propensity score weighting', *Journal of Applied Econometrics* **35**(7), 814–840. 5

Imai, K., Keele, L. & Yamamoto, T. (2010), 'Identification, inference and sensitivity analysis for causal mediation effects', *Statistical Science* pp. 51–71. 2, 3, 4, 5, 6, 7, 8, 16, 29, 34, 36

Imai, K., Tingley, D. & Yamamoto, T. (2013), 'Experimental designs for identifying causal mechanisms', *Journal of the Royal Statistical Society Series A: Statistics in Society* **176**(1), 5–51. 4

Imbens, G. & Angrist, J. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475. 5, 15, 17

Imbens, G. W. (2007), 'Nonadditive models with endogenous regressors', *Econometric Society Monographs* **43**, 17. 4

Imbens, G. W. & Newey, W. K. (2009), 'Identification and estimation of triangular simultaneous equations models without additivity', *Econometrica* **77**(5), 1481–1512. 18, 19

Keele, L., Tingley, D. & Yamamoto, T. (2015), 'Identifying mechanisms behind policy interventions via causal mediation analysis', *Journal of Policy Analysis and Management* **34**(4), 937–963. 10

Kline, P. & Walters, C. R. (2019), 'On heckits, late, and numerical equivalence', *Econometrica* **87**(2), 677–696. 4, 15

Kwon, S. & Roth, J. (2024), 'Testing mechanisms', *arXiv preprint arXiv:2404.11739* . 3

Ludwig, J., Kling, J. R. & Mullainathan, S. (2011), 'Mechanism experiments and policy evaluations', *Journal of economic Perspectives* **25**(3), 17–38. 4

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. 36

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

Roy, A. D. (1951), 'Some thoughts on the distribution of earnings', *Oxford economic papers* **3**(2), 135–146. 3, 13

Słoczyński, T. (2022), 'Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights', *Review of Economics and Statistics* **104**(3), 501–509. 8

Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. (2014), 'Mediation: R package for causal mediation analysis', *Journal of statistical software* **59**, 1–38. https://doi.org/10.18637/jss.v059.i05. 36

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**(1), 331–341. 4, 18

Wang, W. & Yan, J. (2021), 'Shape-restricted regression splines with r package splines2.', *Journal of Data Science* **19**(3). 36

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019), 'Welcome to the tidyverse', *Journal of Open Source Software* **4**(43), 1686. https://doi.org/10.21105/joss.01686. 36

A Appendix

Any comments or suggestions may be sent to me at seh325@cornell.edu, or raised as an issue on the Github project.

A.1 Identification in Causal Mediation

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

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

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

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

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

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

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

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

A.2 Bias in Mediation Estimates

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

A.2.1 Bias in Direct Effect Estimates

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

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

And so,

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

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

To reach the final term, note the following.

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

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

Collect everything together, as follows.

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

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

A.2.2 Bias in Indirect Effect Estimates

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

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

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

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

For each $z' = 0, 1$.

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

So compose the CM estimand, 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 terms between the groups $D_i = 1$ and $D_i = 0$.

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

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

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

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

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

Collect everything together, as follows.

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

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

A.3 A Regression Framework for Direct and Indirect Effects

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

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

$U_{0,i}, U_{1,i}$ are error terms with unknown distributions, mean independent of $Z_i, \boldsymbol{X}_i$ by definition — but possibly correlated with D_i.

Z_i is conditionally independent of potential outcomes, so that $U_{0,i}, U_{1,i} \perp\!\!\!\perp Z_i$. Thus, the first-stage regression of $Z \to Y$ has unbiased estimates.

$$\begin{aligned}
D_i &= Z_i D_i(1) + (1 - Z_i) D_i(0) \\
&= D_i(0) + Z_i \left[D_i(1) - D_i(0)\right] \\
&= \underbrace{\mathbb{E}\left[D_i(0) \,|\, \boldsymbol{X}_i\right]}_{\text{Intercept}} + \underbrace{Z_i \mathbb{E}\left[D_i(1) - D_i(0)\right]}_{\text{Regressor}} \\
&\quad + \underbrace{D_i(0) - \mathbb{E}\left[D_i(0) \,|\, \boldsymbol{X}_i\right] + Z_i\Big(D_i(1) - D_i(0) - \mathbb{E}\left[D_i(1) - D_i(0) \,|\, \boldsymbol{X}_i\right]\Big)}_{\text{Mean-zero independent error term, since } Z_i \perp\!\!\!\perp D_i \,|\, \boldsymbol{X}_i} \\
&=: \phi + \overline{\pi} Z_i + \zeta(\boldsymbol{X}_i) + \eta_i
\end{aligned}$$

$$\implies \mathbb{E}\left[D_i \,|\, Z_i, \boldsymbol{X}_i\right] = \phi + \overline{\pi} Z_i + \zeta(\boldsymbol{X}_i), \text{ and thus unbiased estimates since } Z_i \perp\!\!\!\perp \phi, \eta_i.$$

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.

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

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

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

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

With the following definitions:

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

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

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

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

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

This sequence gives us the resulting regression equation:

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

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

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

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

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

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

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

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

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

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

A.4 Roy Model and Sequential Ignorability

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

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

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

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

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

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.

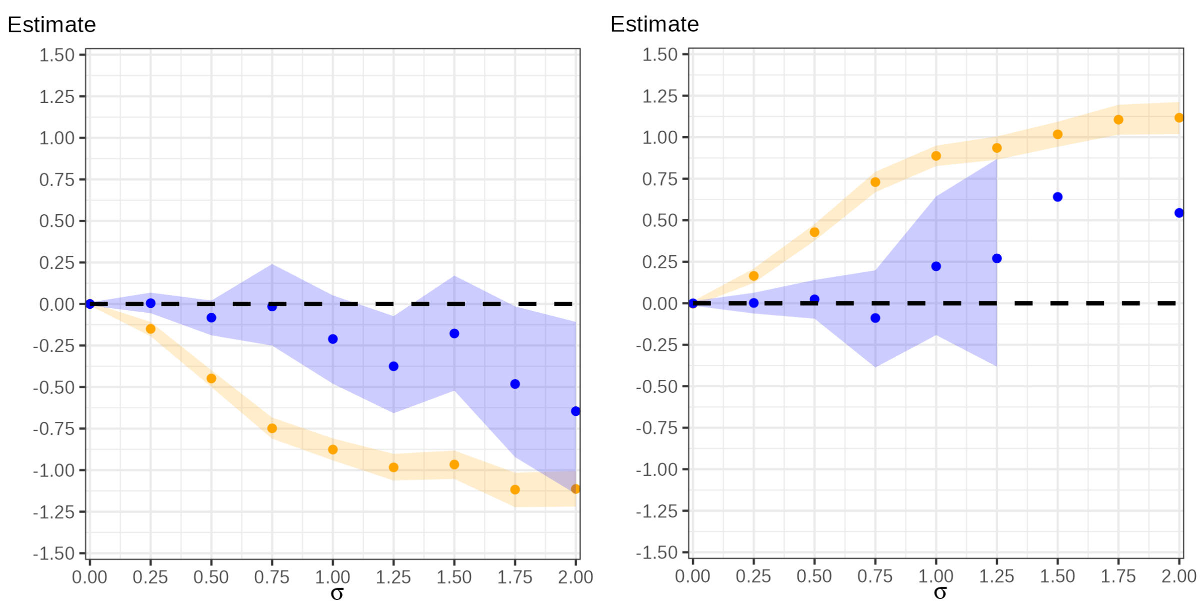
A.5 Control Function Simulation

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

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

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

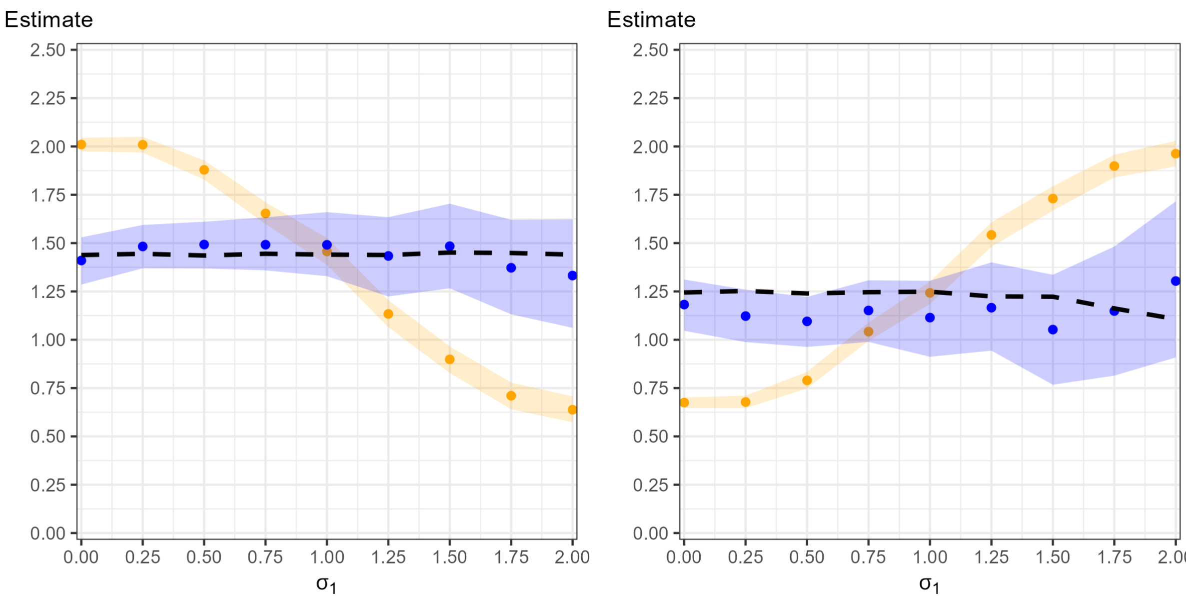
(a) $\widehat{\text{ADE}} - \text{ADE}$. **(b)** $\widehat{\text{AIE}} - \text{AIE}$.



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

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

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



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