# Causal Mediation in Natural Experiments

Senan Hogan-Hennessy*
Economics Department, Cornell University†

This version: 19 January 2025

***Unfinished, please do not circulate.***

**Abstract**

Natural experiments are a cornerstone of applied economics, providing settings for estimating causal effects with a compelling argument for treatment ignorability. Economists are often interested in understanding the mechanisms through which causal treatment effects operate, and Causal Mediation (CM) methods aid this by estimating how much of the treatment effect operates through a proposed mediator. The most popular CM approach relies on assumptions which are unrealistic in natural experiment settings: assuming the mediator is conditionally ignorable — in addition to the ignorability argument for the initial treatment. This paper shows that this approach leads to biased inference, solving for explicit bias terms when the mediator is not ignorable. Using the case of a Roy model for a mediator, I show that individuals' selection based on expected gains and costs is inconsistent with mediator ignorability without implausible behavioural assumptions, and that bias terms are large in practice. I show a control function approach, which overcomes these hurdles if monotonicity holds, using cost of mediator take-up as an instrument. Simulations confirm that this method corrects for persistent bias in conventional CM estimates, and performs comparably to a selection-on-observables approach when the structural assumptions do not hold. This approach gives applied researchers a practical method to estimate CM effects when they can only establish a credible argument for randomisation of the initial treatment, as is common in natural experiments.

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

*For helpful comments I thank Neil Cholli, Lukáš Lafférs, Hyewon Kim, Yiqi Liu, Douglas Miller, Zhuan Pei, Brenda Prallon, and Evan Riehl. Some results in this paper previously circulated in an unpublished version of the working paper "The Direct and Indirect Effects of Genetics and Education." Any comments or suggestions may be sent to me at seh325@cornell.edu.
†Address: Uris Hall #447, Economics Department, Cornell University NY 14853 USA.

Economists use natural experiments to credibly answer social questions, without the trouble of guiding randomisation of what they study. Did Vietnam-era military service lead to income losses? Does access to health insurance lead to employment gains? Do transfer payment lead to measurable long-run economic gains? Quasi-experimental variation gives methods to answer these questions, but give no indication of how these effects came about. Causal Mediation (CM) aims to estimate the mechanisms behind causal treatment 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 large bias terms are in the real world, and an approach to correctly estimate CM effects under minimal structural assumptions.

This paper starts by answering the following question: what does a selection-on-observables CM approach actually estimate when the mediator is not ignorable? Estimates for the average direct and indirect effects are contaminated by bias terms — a sum of selection bias and non-parametric group differences. I then show how this bias operates in an applied regression framework, with bias coming from a correlated error term, showing that the bias term grows larger with the degree of unexplained selection. If individuals have been choosing whether to partake in a mediator based on expected costs and benefits (i.e., following a rational maximisation process), then assuming the mediator is ignorable gives unlikely implications for choice behaviour. This means the identifying assumption for conventional CM methods are unlikely to hold, and likely lead to biased inference in natural experiment settings.

I consider an alternative control function approach to estimating mediation effects. This approach solves the identification problem by instead placing a structural assumption for selection into the mediator (monotonicity), and assumes the researcher has a valid instrument for mediator take-up. These assumptions may hold in real-world natural experiment 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. The existence of a valid instrument is a stronger assumption, which will not hold in every applied example, though is important to avoid parametric assumptions. The most compelling example is using data on the cost of mediator take-up as a first-stage instrument, if it varies between individuals for exogenous reasons and is strong in explaining compliance. Using an instrument avoids parametric assumptions on unexplained mediator selection, though limits the wider applicability of the method. This approach is not perfect: it provides no harbour for estimating CM effects if these structural assumptions do not hold true, though performs no worse than conventional CM methods in this case.

The most popular approach to CM estimates direct and indirect effects by assuming that a treatment is ignorable, and then assuming that a mediator is ignorable conditional on the treatment assignment (Imai et al. 2010). This approach arose in the statistics literature, and is widely used in epidemiology, medicine, and psychology to estimate mediation effects in observational studies. The applied economics literature has not picked up this practice, partially in an understanding that these assumptions are invalid in most observational settings. Indeed, a new strand of the econometric literature has developed estimators for CM effects under overlapping quasi-experimental research designs (Deuchert et al. 2019, Frölich & Huber 2017), a partial identification approach (Flores & Flores-Lagunes 2009), or testing 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 in many cases. This paper makes this part explicit, showing exactly how a conventional approach to CM in a natural experiment can fail in practice.

This paper considers the case when it is not credible to assume the mediator is ignorable (e.g., none of the research designs above apply), leveraging classic labour economic theory for selection-into-treatment to identify direct and indirect effects. A selection-on-observables approach to CM in this setting suffers from bias of the same flavour as classic selection bias (Heckman et al. 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 et al. 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 economic work, which also uses the Roy model as a benchmark (Heckman 1979, Heckman & Honore 1990). I follow the lead of these papers by using a control function approach to correct for the bias developed above. This approach assumes mediator monotonicity, to ensure the mediator follows a selection model (Vytlacil 2002), and a valid instrument for mediator take-up, to avoid parametric assumptions on unobserved selection (Heckman & Navarro-Lozano 2004). Doing so is as an extension of using instruments to identify CM effects — already noted by (Frölich & Huber 2017).[2] Using a control function to estimate CM effects builds on the influential Imai et al. (2010) approach, marrying the CM literature with labour economic theory on selection-into-treatment for the first time.

This paper proceeds as follows. Section 1 introduces CM, and develops expressions for the bias in mediation estimates in natural experiments. Section 2 describes this bias in applied settings with (1) a regression framework, (2) a setting with selection based on costs and benefits. Section 3 solves the identification problem with a control function, assuming a mediator follows a selection model and a researcher observes exogenous variation in cost of

[1] An alternative method to estimate CM effects is ensuring sequential ignorability holds by a running randomised controlled trial for both treatment and mediator at the same time. This setting has been considered in the literature previously, in theory (Imai et al. 2013) and in practice (Ludwig et al. 2011).

[2] Indeed, this paper does not improve on control function methods in any way, 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.

mediator take-up, and gives simulation evidence. Section 4 concludes.

# 1 Direct and Indirect Effects

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

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

Assume $Z_i$ is ignorable.[4]

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

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

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

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

It 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

[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 $Z_i, D_i, Y_i$ to stick to the instrumental variables notation Angrist et al. (1996), more familiar in empirical economics (Angrist & Pischke 2009).

[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.

(and its conditional variant) holds through-out to simplify notation.[5]

2. The reduced-form effect refers to the effect of the treatment on outcome, $Z \to Y$, and is also known as the intent-to-treat effect in experimental settings, or total effect in causal mediation literature.

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

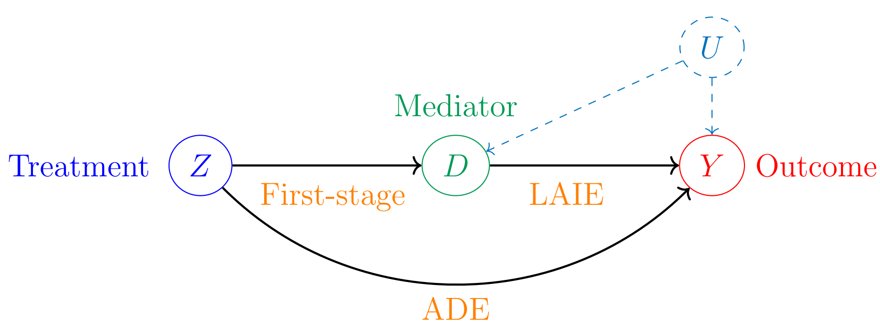
In this setting, $Z_i$ affects outcome $Y_i$ directly, and indirectly via the $D_i(Z_i)$ channel, with no reverse causality. Figure 1 visualises the design, where the direction arrows denote the causal direction (and no reverse causality). On the other hand, mediation aims to decompose the reduced form effect of $Z \to Y$ into these two separate pathways.

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

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

These effects are not separately identified without further assumptions.

**Figure 1:** Structural Causal Model for Causal Mediation.




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

[5]Assuming monotonicity also brings closer to the IV notation, and has other beneficial implications in this setting (see Section 3).

## 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 $\boldsymbol{X}_i$.

**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 assigned randomly, conditional on $\boldsymbol{X}_i$. It then also assumes that, after $Z_i$ is assigned, that $D_i$ is assigned randomly conditional $\boldsymbol{X}_i, Z_i$. If sequential ignorability, 1(1) and 1(2), holds then the direct and indirect effects are identified by two-stage mean differences, after conditioning on $\boldsymbol{X}_i$.[6]

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

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

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 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

[6] Imai et al. (2010) show a general identification statement; I show identification in terms of two-stage regression, 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.

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 in practice, regardless of estimation method.

## 1.2 Bias in Causal Mediation Estimates

Applied research may use a natural experiment to justify the 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 using conventional CM methods to estimate direct and indirect effects, and learn about the mechanisms behind the treatment effect 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 difference terms.*

*Proof.* See Subsection A.2 for the extended 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.

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

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

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

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

The selection bias terms come from systematic differences between the treated and untreated groups, differences not fully unexplained by $\boldsymbol{X}_i$. These selection bias terms would equal to zero if the mediator was ignorable 1(2), but do not necessarily average to zero if not. The group differences represent the fact that a matching estimator 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).[7,8] The group differences term is a non-parametric framing of the bias from controlling for intermediate outcomes, previously studied only in a linear setting. These are referred to as bad controls by Cinelli et al. (2024), or M-bias by Ding & Miratrix (2015).

[7]The group differences term is longer for the AIE estimate, because the indirect effect is comprised from the effect of $D_i$ local to $Z_i$ compliers; a matching estimator gets the average effect on treated, and the longer term adjusts for differences with the complier average effect.

[8]The selection-on-observables approach could, instead, focus on the average effect on treated populations (as do Keele et al. 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 local to the $D_i = 1$ group for the AIE. 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.

# 2 Causal Mediation in Applied Settings

In this section, I further develop the issue of selection in causal mediation 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.

## 2.1 Regression Framework

Inference for direct and direct effects can be written in a regression framework, showing how correlation between the error term and the mediator persistently biases estimates. Write $Y_i(Z, D)$ as a sum of observed factors $Z_i, \boldsymbol{X}_i$ and unobserved factors, $U_{1,i}, U_{0,i}$ (following the notation of Heckman & Vytlacil 2005). Put $\mu_D(Z; \boldsymbol{X}_i) = \mathbb{E}\left[Y_i(Z_i, 0) \,|\, \boldsymbol{X}\right]$, to give a representation of the average direct and indirect effects.

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

Then define the error terms.

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

With this notation, observed data $Z_i, D_i, Y_i$ take the following representation, which characterises direct effects, indirect effects, and bias from selection.

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

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

First-stage (3) is identified, with $\phi, \varphi(\boldsymbol{X}_i)$ the intercept, and $\pi$ the average rate of compliance (which may depend on $\boldsymbol{X}_i$). Second-stage (4) is not identified without further assumptions. $\alpha, \zeta(\boldsymbol{X}_i)$ are the intercept terms, and $\beta, \gamma, \delta$ are values that comprise mediation effects — all whose values may depend on $\boldsymbol{X}_i$, see Subsection A.3 for full definitions. $U_{0,i} + D_i \left(U_{1,i} - U_{0,i}\right)$ is the possibly correlated error term, which disrupts identification. The average direct and indirect effects are averages of these coefficients.

$$\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] = \mathbb{E}\left[\gamma + \delta D_i\right],$$
$$\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] = \mathbb{E}\left[\pi\left(\beta + \delta Z_i\right)\right].$$

By construction, $U_i = U_{1,i} - U_{0,i}$ is an unobserved confounder. The regression estimates of second-stage (4) give unbiased estimates only if $D_i$ is also conditionally ignorable: $D_i \perp\!\!\!\perp U_i$. If not, then regression estimates suffer from omitted variables bias if they do not adjust for the unobserved confounder $U_i$.

## 2.2 Selection on Costs and Benefits

The key to noting that CM is at risk of bias is noting that $D_i \perp\!\!\!\perp U_i$ is unlikely to hold in applied settings. Without an identification strategy for $D_i$, in addition to one for $Z_i$, 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, after $Z_i, \boldsymbol{X}_i$ have been assigned. Write $C_i$ for individual $i$'s costs of taking mediator $D_i$, and $\mathbb{1}\{.\}$ for the indicator function. The Roy model has $i$ taking the mediator if the benefits exceed the costs.

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

Paragraph here talking about why the Roy model is useful. (Roy 1951, Heckman & Honore 1990).

Decompose the costs into its mean and unobserved error, as above $C_i(Z_i) = \mu_C(Z_i; \boldsymbol{X}_i) + U_{C,i}$, and collect the mean benefits, $\mu \coloneqq \mu_1 - \mu_0$. So we can write the first-stage selection equation separated by observed means and unobserved errors.

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

Theorem: if selection is Roy style, and sequential ignorability holds, then unobserved benefits play no part in selection. The only driver in differences in selection are differences in costs (and not benefits).

$$\mathbb{E}\left[D_i(z') \,|\, U_i = u\right] = \mathbb{E}\left[D_i(z') \,|\, U_i = u'\right]$$

For all $u', u$ in the range of the distribution of $U_i$. Proof: by contradiction, add to the appendix. This could, for example, hold if $U_{1,i} - U_{0,i}$ is degenerate conditional on $\boldsymbol{X}_i$.

Short paragraph on why this means $\boldsymbol{X}_i$ must be incredibly rich. Mention the Laffers result. Write about if $D_i$ is the choice to attend education, then $\boldsymbol{X}_i$ must soak up all gains to education. Or assuming that all variation in $D_i$ comes from unobserved differences in take-up costs. This is unlikely to hold true, absent a separate research design for $D_i$, limiting the selection to an information restricted version of the Roy model.

If not, then selection bias propagates, including writing here for what the selection bias term is equal to.

# 3 Solving Identification with a Control Function

If your goal is to estimate CM effects, and you could control for unobserved selection into $D_i$, then you would. The control function method takes this insight seriously, and provides minimal conditions to model and control for the implied unobserved confounder, $U_i$.[9]

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

**Definition 2.** *Control function assumptions.*

$$\Pr\left(D_i(1) \geq D_i(0) \,|\, \boldsymbol{X}_i\right) = 1 \tag{5}$$

$$\boldsymbol{X}_i^{IV} \textit{ has the property } \frac{\partial \mu(\boldsymbol{X}_i)}{\partial \boldsymbol{X}_i^{IV}} = 0 < \frac{\partial D_i(z')}{\partial \boldsymbol{X}_i^{IV}}, \text{ for } z' = 0, 1. \tag{6}$$

Assumption 2(5) is the (conditional) monotonicity assumption (Imbens & Angrist 1994), which is untestable but acceptable in many empirical applications. Assumption 2(6) is assuming that an instrument exists, which satisfies an exclusion restriction (i.e., not impacting mediator gains $\mu$), 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 IV literature.

Write $K_i$ for the mediator propensity score, as a function of the instrument $\boldsymbol{X}_i^{\text{IV}}$ and remaining controls $\boldsymbol{X}_i^-$. $K_i$ serves as the control function in this setting.

$$K_i = \Pr\left(D_i = 1 \,\middle|\, Z_i, \boldsymbol{X}_i^{\text{IV}}, \boldsymbol{X}_i^-\right)$$

**Theorem 2.** *If 2(5) and 2(6) hold, then the average potential outcomes (and thus, the ADE*

[9]This section does not improve on the control function approach, instead only noting its utility to solve the identification problem of CM in a natural experiment setting.

*and AIE) are identified by a control function approach .*

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

*Proof.* Special case of Imbens & Newey (2009, Theorem 1); see Subsection A.4. □

Assumption 2(5) guarantees that mediator $D_i(.)$ can be represented by a selection model (Vytlacil 2002), and 2(6) pins down a control function to identify the selection model.

## 3.1 Relationship to the Roy Model

Writing here about how the Roy model is solves with this, and the instrument impacts costs of mediator take-up.

## 3.2 Estimation and Simulation Evidence

Write first about how to estimate this practically, focus on OLS.

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

$$Z_i \sim \text{Binom}\left(\frac{1}{2}\right), \quad \boldsymbol{X}_i^- \sim N(0,1), \quad \boldsymbol{X}_i^{\text{IV}} \sim \text{Binom}\left(\frac{1}{2}\right),$$
$$\begin{bmatrix} U_{1,i} \\ U_{0,i} \end{bmatrix} \sim \text{BivariateNormal}\left(\begin{bmatrix} 0, 0 \end{bmatrix}, \begin{bmatrix} \sigma^2, \rho\sigma \\ \rho, \sigma^2 \end{bmatrix}\right), \quad U_{C,i} = 0.$$

Additionally, suppose each $i$ chooses to take mediator $D_i$ based on the costs and benefits (i.e., a Roy model), with following definitions for each $z', d' = 0, 1$.

$$\mu_{d'}\left(z'; \boldsymbol{X}_i\right) = \boldsymbol{X}_i^{-} + \left(z' + d' + z'd'\right), \quad \mu_C\left(z'; \boldsymbol{X}_i\right) = \frac{z'}{2} + \boldsymbol{X}_i^{-} \boldsymbol{X}_i^{\mathrm{IV}}.$$

Following Section 2, these data have the following outcome equations:

$$D_i = \mathbb{1}\left\{ \frac{Z_i}{2} - \boldsymbol{X}_i^{-} \boldsymbol{X}_i^{\mathrm{IV}} \geq U_{0,i} - U_{1,i} \right\},$$
$$Y_i = D_i + Z_i + Z_i D_i + \boldsymbol{X}_i^{-} + \left(1 - D_i\right) U_{0,i} + D_i U_{1,i}.$$

In this setting the error terms $U_{i,0}, U_{i,1}$ determine the bias in OLS estimates of the ADE and AIE, so the bias varies for different values of the DGP parameters $\rho \in [-1, 1]$ and $\sigma_0 = 1, \frac{\sigma_1}{\sigma_0} = \sigma$.

**To-do:** write the formula for main confounders, $\mathbb{E}\left[U_{0,i}\right] D_i = 0$ and $\mathbb{E}\left[U_{1,i}\right] D_i = 1$, which are the bias terms, and how they depend on $\rho, \sigma_0, \sigma_1$.

# 4 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 how individuals select into as a way of judging CM methods. 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. 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 (Huber 2020, Bia et al. 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; if they 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 settings with both data and 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. 7

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. 4

Angrist, J. D. & Pischke, J.-S. (2009), *Mostly harmless econometrics: An empiricist's companion*, Princeton university press. 4

Athey, S., Tibshirani, J. & Wager, S. (2019), 'Generalized random forests', *The Annals of Statistics* **47**(2), 1148–1178. 18

Bach, P., Chernozhukov, V., Kurz, M. S., Spindler, M. & Klaassen, S. (2024), 'DoubleML — An object-oriented implementation of double machine learning in R'. `https://doi.org/10.18637/jss.v108.i03`. 18

Bia, M., Huber, M. & Lafférs, L. (2024), 'Double machine learning for sample selection models', *Journal of Business & Economic Statistics* **42**(3), 958–969. 15

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

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

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

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

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

Heckman, J., Ichimura, H., Smith, J. & Todd, P. (1998), 'Characterizing selection bias using experimental data', *Econometrica* **66**(5), 1017–1098. 3, 8

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

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

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

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

Hlavac, M. (2018), *stargazer: Well-Formatted Regression and Summary Statistics Tables*, Central European Labour Studies Institute (CELSI). R package version 5.2.2, `https://CRAN.R-project.org/package=stargazer`. 18

Huber, M. (2020), 'Mediation analysis', *Handbook of labor, human resources and population economics* pp. 1–38. 2, 15

Imai, K., Keele, L. & Yamamoto, T. (2010), 'Identification, inference and sensitivity analysis for causal mediation effects', *Statistical Science* pp. 51–71. 2, 3, 4, 6, 18, 23

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

Imbens, G. & Angrist, J. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475. 4, 12

Imbens, G. W. (2007), 'Nonadditive models with endogenous regressors', *Econometric Society Monographs* **43**, 17. 3

Imbens, G. W. & Newey, W. K. (2009), 'Identification and estimation of triangular simultaneous equations models without additivity', *Econometrica* **77**(5), 1481–1512. 13

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. 8

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

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

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`. 18

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

Roy, A. D. (1951), 'Some thoughts on the distribution of earnings', *Oxford economic papers* **3**(2), 135–146. 3, 11

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. 7

Tibshirani, J., Athey, S., Sverdrup, E. & Wager, S. (2023), *grf: Generalized Random Forests*. R package version 2.3.0, `https://CRAN.R-project.org/package=grf`. 18

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**(1), 331–341. 3, 13, 24

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`. 18

# A Appendix

This project used computational tools which are fully open-source. Any comments or suggestions may be sent to me at seh325@cornell.edu, or raised as an issue on the Github project.

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

- *Tidyverse* (Wickham et al. 2019) collected tools for data analysis in the R language.
- *DoubleML* (Bach et al. 2024) implemented doubly robust methods used in the empirical analysis.
- *GRF* (Athey et al. 2019, Tibshirani et al. 2023) compiled forest computational tools for the R language.
- *Stargazer* (Hlavac 2018) provided methods to efficiently convert empirical results into presentable output in LaTeX.

## A.1 Identification in Causal Mediation

Imai et al. (2010, Theorem 1) states that the direct and indirect effects 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 average direct effect is a simple rearrangement of the above. The estimand for the average indirect effect relies on a different sequence, relying on (1) sequential ignorability, (2) conditional monotonicity. These give (1) identification of, and equivalence between, LADE conditional on $\boldsymbol{X}_i$ and ADE conditional on $\boldsymbol{X}_i$, (2) identification of the complier score.

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

Monotonicity is not technically required for the above. Breaking monotonicity would not change the identification of any of the above; it would be the same except replacing the complier score with a complier or 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 terms 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 a difference term between the average and the average for relevant complier groups.

Collect everything together, as follows.

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

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

#### A.2.2 Bias in Indirect Effect Estimates

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

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

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

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

For each $z' = 0, 1$.

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

So compose the CM estimand, as follows.

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

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

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

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

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

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.

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

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 $\mu_D(Z; \boldsymbol{X}) = \mathbb{E}\left[Y_i(Z, D) \mid \boldsymbol{X}\right]$ and $U_{D,i} = Y_i(Z, D) - \mu_D(Z; \boldsymbol{X})$, so we have the following expressions.

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

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

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

$\implies \mathbb{E}\left[D_i \mid Z_i, \boldsymbol{X}_i\right] = \phi + \pi Z_i + \varphi(\boldsymbol{X}_i)$, 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.

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

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

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

With the following definitions:

**(a)** $\alpha = \mathbb{E}\left[\mu_0(0; \boldsymbol{X}_i)\right]$ and $\zeta(\boldsymbol{X}_i) = \mu_0(0; \boldsymbol{X}_i) - \alpha$ 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)** $\delta = \mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - \left(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\right)$ is the interaction effect.

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

This sequence gives us the resulting regression equation:

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

Taking the conditional expectation, and collecting for the expressions of the direct and indirect effects:[10]

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

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.

[10] 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.

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

And so the direct and indirect effect estimates are contaminated by these bias terms.

## A.4 Control Function Identification

Write the proof in here, following Vytlacil (2002) construction in the forward direction. Note that the notation needs updating for no exclusion restriction.

And then