Causal Mediation in Natural Experiments

Senan Hogan-Hennessy*
Economics Department, Cornell University†

This version: 18 December 2024

Aspirational Abstract, please do not circulate.

Abstract

Natural experiments are a cornerstone of applied economics, providing settings for estimating causal effects of a treatment with a compelling argument for treatment ignorability. Economists are often interested in understanding the mechanisms through which causal effects operate, and mediation methods aim to estimate these components. However, conventional mediation methods rely on a selection-on-observables assumption, assuming the mediator is conditionally ignorable — in addition to the natural experiment for the original treatment. This paper shows that conventional estimates of mediation effects are contaminated by selection bias 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. I develop a control function approach, which correctly estimates mediation effects when selection into the mediator follows a selection model, using cost of mediator take-up as an instrument. Simulations confirm that this method corrects for selection bias in conventional mediation estimates, and performs comparably to a selection-on-observables approach when the mediator selection does not follow a selection model. I illustrate the approach by estimating the proportion of the causal effect of genes associated with education that operates via a direct genetic channel versus indirectly through extended schooling. Finally, I provide an implementation of this method in the *R* package *mediate-controlfun*, offering an accessible tool for robust mediation analysis in natural experiment settings.

Keywords: Causal Mediation.

JEL Codes: D31, D91, I24, J24, Z00.

*For helpful comments I thank Neil Cholli, Lukáš Laffers, Yiqi Liu, Douglas Miller, Zhuan Pei, and Brenda Prallon. Any comments or suggestions may be sent to me at seh325@cornell.edu.

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

1 Introduction

Conventional CM methods rely on a selection-on-observables assumption, which may not hold true in observational work. I explicitly connect the assumptions behind CM methods to those of selection into treatment in classic labour and observation economic research (Heckman & Vytlacil 2005). When a mediator, here education, is not randomly assigned then conventional CM methods for estimating direct and indirect effects are contaminated by selection bias. I write this as both a non-parametric non-identification result, and with a model-based regression framework with a correlated error term(e.g., as in the Imai et al. 2010 linear model approach). Structural assumptions could solve the identification problem, for example if selection into education follows a Roy model or errors terms have a known distribution (Heckman 1979). Adjusting indirect and direct estimates with sample selection models gives estimates for the direct genetic channel indistinguishable from zero, assigning roughly all the association for Ed PGI to earnings by the education channel.

This work adds to a growing literature of genetics in economics (Barth et al. 2020), and expands on mediation methods (Imai et al. 2010) which are rarely used in empirical economic research (Huber 2020). The most similar papers have studied the association between Ed PGI and earnings (Papageorge & Thom 2020), and socioeconomic status (Carvalho 2024). Another has considered a similar topic from the view of genes as instruments, when the exclusion restriction is violated (Spiller et al. 2019, Van Kippersluis & Rietveld 2018). To the best of my knowledge, this is the first paper to connect mediation methods (and its selection-on-observables assumptions) to the labour economics literature for selection into treatment.

2 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, and Y_i an outcome for individuals $i = 1, \ldots, n$. The outcomes are a sum of their potential outcomes:

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

Z_i affects outcome Y_i directly, and indirectly via the $D_i(Z_i)$ channel, with no reverse causality. The framework is general to any (conditionally) randomly assigned Z_i, selected mediator D_i, and outcome Y_i. For the analysis in **??**, Z_i is the Ed PGI, D_i a measure of education year, and Y_i (log) later-life earnings. Note that Z_i, D_i are continuous measures in HRS Data, but this section focuses on binary $Z_i, D_i = 0, 1$ to simplify the causal framework.[1] Figure 1 visualises the design, where the direction arrows denote the causal direction (and no reverse causality).

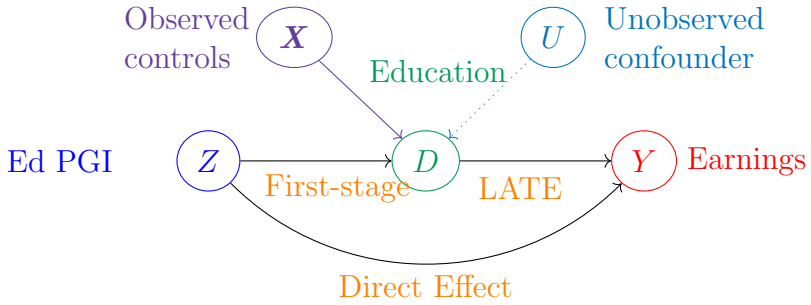
Assuming that Ed PGI Z_i is randomly assigned (or conditionally so, see **??**), then there are only two average effects which are identified. The first-stage effect refers to the effect of the Ed PGI on education, $Z \to D$.

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

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

[1] Continuous analogues of the following are extensions of the binary case, and will be included in work to follow.

Figure 1: Structural Causal Model for Direct and Indirect Effects of Genetics and Education.



Note: This figures shows the structural causal model for decomposing the direct and indirect effects of genetics and education attainment.

(and its conditional variant) holds through-out, as it brings the mediation notation closer to the IV literature from labour economics.[2]

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

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

On the other hand, mediation aims to decompose the reduced form effect of $Z \to Y$ into two separate pathways: indirectly through D, and directly absent D.

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

These effects are not separately identified without further assumptions.

[2]Monotonicity has other beneficial implications in this setting, as shown in Subsection 3.1.

2.1 Causal Mediation Estimates

The conventional approach to estimating direct and indirect effects assumes both Z_i and D_i are conditionally ignorable.

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

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

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

If 1(1) and 1(2) hold, then the direct and indirect effects are identified by two-stage mean differences, after conditioning on $\boldsymbol{X}_i$:[3]

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

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

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

These estimands are typically estimated with linear models (Imai et al. 2010):

$$D_i = \phi + \pi Z_i + \boldsymbol{\psi}_1' \boldsymbol{X}_i + \eta_i$$

$$Y_i = \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \boldsymbol{\psi}_2' \boldsymbol{X}_i + \varepsilon_i$$

And so the direct and indirect effects are composed from OLS estimates, $\widehat{\gamma} + \widehat{\delta}\mathbb{E}\left[D_i\right]$ for the direct effect and $\widehat{\pi}\left(\widehat{\beta} + \mathbb{E}\left[Z_i\right]\widehat{\delta}\right)$ for the indirect effect. While this is the most common

[3] 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.

approach in the applied literature, I do not focus on the linear formulation of this problem as it assumes homogenous treatment effects and linear confounding. These assumptions are unnecessary to my analysis; it suffices to note that heterogeneous treatment effects and non-linear confounding would bias OLS estimates of direct and indirect effects in the same manner that is well documented elsewhere (see e.g., Angrist 1998, Słoczyński 2022). I focus on fundamental problems that plague causal mediation methods in practice, regardless of estimation method. As such, I focus my work on non-parametric identification, and employ semi- and non-parametric estimation methods in my empirical analysis whenever possible to avoid these problems.

2.2 Selection Bias in Causal Mediation Estimates

Mediation methods are the main method that researchers then answer the following question: how did Z lead to a causal effect on Y, and through which channels? In observational work this may include a natural experiment that quasi-randomly assigns Z_i to individuals, regardless of their preferences or selection patterns — i.e., justifying assumption 1(1). Rarely does observational research employ an additional, overlapping identification design for D_i as part of the analysis, and instead they employ mediation methods by assuming this D_i is ignorable given observed covariates $\boldsymbol{X}_i$.[4] This approach leads to biased estimates, and contaminates inference regarding direct and indirect effects (in the same manner as Heckman et al. 1998).

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 causal mediation estimates are contaminated by selection bias terms, and group differences terms.*

Proof. See Subsection A.4 for the extended proof. □

[4] Imai et al. (2013) call attention to the need for a separate research design to isolate causal effects of D_i in randomised controlled trials; Subsection A.3 overviews literature, finding many papers that employ mediation methods with a research design for Z_i, but not for D_i.

Below I present the relevant selection bias and group difference terms, omitting the conditional on $\boldsymbol{X}_i$ notation for brevity. These selection bias terms would be equal to zero if the mediator was conditional ignorable (2), but do not necessarily average to zero if not.

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

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

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

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

The selection bias terms come from systematic differences between the treated and untreated groups, differences not fully unexplained by $\boldsymbol{X}_i$. 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).[5] The group differences term is longer for the average indirect effect 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.

3 Direct and Indirect Effects Under Selection

This section connects causal mediation, without assuming the mediator is randomly assigned (i.e., under selection), with classic labour economics models for selection into treatment.

3.1 Selection Model Representation

The IV literature assumes a first-stage monotonicity condition, where randomised Z_i influences mediator D_i in at most one direction.

Definition 2. *First-stage Monotonicity (Imbens & Angrist 1994).*

$$\Pr\left(D_i(1) \geq D_i(0)\right) = 1 \tag{3}$$

Assuming 2(3) in a mediation setting opens mediation to the wide literature on IV and selection models for identification in the presence of selection.

Theorem 2. *Under monotonicity, mediator D_i can be represented by a selection model. Suppose 2(3) holds, then there is a function $\mu(.)$ and random variable U_i such that D_i takes the following form.*

$$D_i(z) = \mathbb{1}\left\{\mu(z) \geq U_i\right\}, \quad \forall z = 0, 1$$

[5]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 estimate for the ADE on treated gives the ADE local to the $Z_i = 1$ treated group, while the AIE estimate gives the AIE local to the $D_i = 1$ group. In this way, these ADE and AIE on treated terms are not comparable to each other, so I focus on the true average terms.

Proof. Special case of the Vytlacil (2002) equivalence result; see Subsection A.5. □

Theorem 2 is a powerful result: it says that at the cost of assuming monotonicity (as is done in the IV literature), then selection into D_i takes a latent index form, and opens up identification in a mediation context to the wide literature on identifying treatment effects in selection models.

3.2 A Regression Framework for Direct and Indirect Effects

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. And thus, selection models can be used to adjust for unobserved confounding.

To motivate a regression framework with unobserved confounding, write $Y_i(Z, D)$ as a sum of observed factors $Z_i, \boldsymbol{X}_i$ and unobserved factors, (following the notation of Heckman & Vytlacil 2005). First, define the following unobserved error terms

$$U_{0,i} = Y_i(Z_i, 0) - \mathbb{E}\left[Y_i(Z_i, 0) \mid \boldsymbol{X}\right], \quad U_{1,i} = Y_i(Z_i, 1) - \mathbb{E}\left[Y_i(Z_i, 1) \mid \boldsymbol{X}\right]$$

Then observed data take the following representation, which characterises direct effects, indirect effects, and the selection problem (see Subsection A.6 for all definitions).

$$\begin{aligned} D_i &= \phi_i + \pi_i Z_i + \eta_i \\ Y_i &= \alpha_i + \beta_i D_i + \gamma_i Z_i + \delta_i Z_i D_i + \underbrace{U_{0,i} + D_i\left(U_{1,i} - U_{0,i}\right)}_{\text{Correlated error term.}} \end{aligned}$$

And the average direct and indirect effects are given by the following

$$\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] = \mathbb{E}\left[(\beta_i + Z_i\delta_i) \times \pi_i\right],$$
$$\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] = \mathbb{E}\left[\gamma_i + \delta_i D_i\right].$$

By assumption $Z_i \perp\!\!\!\perp Y_i(.,.), D_i(.)$, so that the regression only gives unbiased estimates if D_i is also conditionally random: $D_i \perp\!\!\!\perp U_{0,i} - U_{1,i} \mid \boldsymbol{X}_i$. If not, then the regression estimates (without adjusting for the contaminated bias term) suffer from omitted variables bias.

3.3 Selection into Education

In the education context, point identifying direct and indirect effects requires the *researcher controls for all sources of selection-into-education.*

While this assumption may hold true in two-way randomised experiments (e.g., in a laboratory or two-way RCT), it is unlikely to hold in the case of quasi-experimental variation in Z, or when modelling education as a mediator — absent a separate identification strategy for education D. To expand this point in an econometric selection-into-treatment framework, suppose selection follows a Roy model, where individual i weighs the costs and benefits of completing education.

$$D_i(Z_i) = \mathbb{1}\left\{\underbrace{C_i(Z_i)}_{\text{Costs}} \leq \underbrace{Y_i(Z_i, 1) - Y_i(Z_i, 0)}_{\text{Gains}}\right\}$$

Education choice $D_i(z)$ is clearly related to $Y_i(z, d)$ in this model, so let's see what the equation looks like in terms of sequential ignorability. As above, decompose costs into observed and unobserved factors.

$$C_i(Z_i) = \mu_C(Z_i; \boldsymbol{X}_i) + U_{C,i}$$

And so we can write the first-stage selection equation in full.

$$D_i(z) = \mathbb{1}\left\{ \underbrace{U_{C,i} + U_{0,i} - U_{1,i}}_{\text{Unobserved}} \leq \underbrace{\mu_1(z; \boldsymbol{X}_i) - \mu_0(z; \boldsymbol{X}_i) - \mu_C(z; \boldsymbol{X}_i)}_{\text{Observed}} \right\}$$

Sequential ignorability, where $Y_i(z, d) \perp\!\!\!\perp D_i(z') \mid \boldsymbol{X}_i$, would then require that $\mathbb{E}\left[U_{0,i} - U_{1,i} \mid D_i\right] = 0$. In the Roy model above, this would assume every single contribution for returns to education is contained in $\boldsymbol{X}_i$; if there are any unobserved sources by which people have systematically different returns to education, then they would select into education based on this, and bias naïve mediation estimates. This is unlikely to hold true, unless there is another identification strategy for D_i — in addition to the one used for Z_i.

3.4 Estimating Direct and Indirect Effects

Quasi-experimental work does not take the assumption of "selection-on-observables" at face value without an explicit research design (Angrist & Pischke 2009) or modelling approach to address this issue.

A classical approach to modelling this issue, is a selection model approach (Heckman 1974, 1979). The approach assume U_0, U_1 follow a known distribution (e.g, bivariate normal), and estimates the regression via maximum likelihood. Alternatively, a control function approach estimates the system in two stages, avoiding (some) distributional assumptions if an instrument is used, at the cost of efficiency. In the following, I estimate direct and indirect effects first by OLS (assuming sequential ignorability), and then via both variants of the sample selection models, to compare estimates. Future work will consider estimates by using an alternative instrument for education, in the framework of Frölich & Huber (2017) to avoid the modelling assumptions inherent to sample selection models.

4 Discussion and Future Work

This project aims to achieve two main goals: first, to test the claims that genetics (specifically Ed PGI) is associated with labour market outcomes independently, and secondly to connect the mediation literature to classical labour economic methods for adjusting for selection bias. Adjusting conventional mediation methods via a structural selection model for education reduces estimates for the direct channel in the genetic association from 50% to 0%. These results bring into question previous claims, in the context of Ed PGI, that genetics affect outcomes independently of education.

This work so far has focused on genetic association, and not causal effects, because the HRS data have no clear research design for random variation in Ed PGI. This means that the above estimates are only correlational because the EA Score is heritable, and not randomized. However, there is opportunity to analyse random genetic variation, thanks to Mendelian independent assortment. If a father has an Ed PGI of X and a mother Y, then genetic mixing at conception means their child is expected to have an EA Score of $\frac{X+Y}{2}$. Thanks to genetic mixing, their child may have EA Score above or below the expected value, as they randomly inherited more/fewer genes in the EA Score from the parent with a higher score (a.k.a. random Mendelian segregation, **?**). The HRS has no data on parental genetic information, so the estimates above did not control for parents' scores and are thus not causal (Young et al. 2022). In-progress work is expanding on the above, using UK Biobank data on genetic data after controlling for parents genes, expanding these results from genetic associations to genetic effects.

Secondly, this project has so far connected causal mediation to classical approaches to selection into treatment, using a Roy model as a key structural example for which selection models can overcome selection bias in mediation analyses. However, an explicit research design for years of education (in addition to Ed PGI) is necessary for realistic estimates — in the sense of a causally identified analysis (Angrist & Pischke 2009). An overlapping

instrument for years of education is necessary to compare to the results of classical sample selection models.

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. 5

Angrist, J. D. & Pischke, J.-S. (2009), *Mostly harmless econometrics: An empiricist's companion*, Princeton university press. 10, 11

Athey, S., Tibshirani, J. & Wager, S. (2019), 'Generalized random forests', *The Annals of Statistics* **47**(2), 1148–1178. 15

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. 15

Barth, D., Papageorge, N. W. & Thom, K. (2020), 'Genetic endowments and wealth inequality', *Journal of Political Economy* **128**(4), 1474–1522. 1

Carvalho, L. S. (2024), 'Genetics and socioeconomic status: Some preliminary evidence on mechanisms'. 1

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. 10

Heckman, J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica: journal of the econometric society* pp. 679–694. 10

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

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica: Journal of the econometric society* pp. 153–161. 1, 10

Heckman, J. J. & Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation 1', *Econometrica* **73**(3), 669–738. 1, 8

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. 15

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

Imai, K., Keele, L. & Yamamoto, T. (2010), 'Identification, inference and sensitivity analysis for causal mediation effects', *Statistical Science* pp. 51–71. 1, 4, 15

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. 5

Imbens, G. & Angrist, J. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475. 2, 7

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. 7

Papageorge, N. W. & Thom, K. (2020), 'Genes, education, and labor market outcomes: evidence from the health and retirement study', *Journal of the European Economic Association* **18**(3), 1351–1399. 1

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/* 15

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. 4

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. 5

Spiller, W., Slichter, D., Bowden, J. & Davey Smith, G. (2019), 'Detecting and correcting for bias in mendelian randomization analyses using gene-by-environment interactions', *International journal of epidemiology* **48**(3), 702–712. 1

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. 15

Van Kippersluis, H. & Rietveld, C. A. (2018), 'Pleiotropy-robust mendelian randomization', *International journal of epidemiology* **47**(4), 1279–1288. 1

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**(1), 331–341. 8, 16

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. 15

Young, A. I., Nehzati, S. M., Benonisdottir, S., Okbay, A., Jayashankar, H., Lee, C., Cesarini, D., Benjamin, D. J., Turley, P. & Kong, A. (2022), 'Mendelian imputation of parental genotypes improves estimates of direct genetic effects', *Nature genetics* **54**(6), 897–905. 11

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 Continuous Average Causal Responses

Section here relating the approach to the average causal response function (see e.g., Angrist Imbens JASA 1996, Andrew Bacon for DiD 2023).

A.3 Previous Literature

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

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

A.4 Selection Bias in Mediation Estimates

Write the proof in here. It is long....

A.5 Proof of the Selection Model Representation

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

A.6 A Regression Framework for Direct and Indirect Effects

Put $\mu_D(Z; \boldsymbol{X}) = \mathbb{E}\left[Y_i(Z, D) \,|\, \boldsymbol{X}\right]$, 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 assumed randomly assigned, independent of potential outcomes, so that $U_{0,i}, U_{1,i} \perp\!\!\!\perp Z_i$. Thus, the first-stage $Z \to Y$ leads to unbiased estimates.

$$\begin{aligned}
D_i &= Z_i D_i(1) + (1 - Z_i) D_i(0) \\
&= D_i(0) + Z_i \left[D_i(1) - D_i(0) \right] \\
&= \underbrace{\mathbb{E}\left[D_i(0) \right]}_{\text{Intercept}} + \underbrace{Z_i \mathbb{E}\left[D_i(1) - D_i(0) \right]}_{\text{Regressor}} \\
&\quad + \underbrace{D_i(0) - \mathbb{E}\left[D_i(0) \right] + Z_i \big(D_i(1) - D_i(0) - \mathbb{E}\left[D_i(1) - D_i(0) \right] \big)}_{\text{Mean-zero, independent error term}} \\
&=: \phi_i + \pi_i Z_i + \eta_i \\
\implies \mathbb{E}\left[D_i \,|\, Z_i \right] &= \mathbb{E}\left[\phi_i \right] + \mathbb{E}\left[\phi_i \right] Z_i + \mathbb{E}\left[\eta_i \right], \text{ and thus unbiased estimates since } Z_i \perp\!\!\!\perp \phi_i, \eta_i.
\end{aligned}$$

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.

$$\begin{aligned}
\mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i \right] &= \mu_0(0; \boldsymbol{X}_i) \\
&\quad + Z_i \left[\mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i) \right] \\
&\quad + D_i \left[\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i) \right] \\
&\quad + Z_i D_i \left[\mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - (\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)) \right] \\
&\quad + U_{0,i} + D_i \left(U_{1,i} - U_{0,i} \right) \\
&=: \alpha_i + \beta_i D_i + \gamma_i Z_i + \delta_i Z_i D_i + \underbrace{U_{0,i} + D_i \left(U_{1,i} - U_{0,i} \right)}_{=\varepsilon_i}
\end{aligned}$$

$\alpha_i, \beta_i, \delta_i$ are the relevant direct effect under $D_i = 1$, indirect effect under $Z_i = 1$, δ_i the interaction effect, and ε_i the remaining error term. Collecting for the expressions of the direct and indirect effects:[6]

$$\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] = \mathbb{E}\left[\pi_i\left(\beta_i + Z_i\delta_i\right)\right]$$
$$\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] = \mathbb{E}\left[\gamma_i + \delta_i D_i\right]$$

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:

$$\mathbb{E}\left[\widehat{\alpha}\right] = \mathbb{E}\left[\alpha_i\right] + \mathbb{E}\left[D_i\left(U_{1,i} - U_{0,i}\right)\right]$$
$$\mathbb{E}\left[\widehat{\beta}\right] = \mathbb{E}\left[\beta_i\right] + \frac{\mathrm{Cov}\left(D_i,\ D_i\left(U_{1,i} - U_{0,i}\right)\right)}{\mathrm{Var}\left(D_i\right)}$$
$$\mathbb{E}\left[\widehat{\gamma}\right] = \mathbb{E}\left[\gamma_i\right] + \frac{\mathrm{Cov}\left(Z_i,\ D_i\left(U_{1,i} - U_{0,i}\right)\right)}{\mathrm{Var}\left(Z_i\right)}$$
$$\mathbb{E}\left[\widehat{\delta}\right] = \mathbb{E}\left[\delta_i\right] + \frac{\mathrm{Cov}\left(Z_iD_i,\ D_i\left(U_{1,i} - U_{0,i}\right)\right)}{\mathrm{Var}\left(Z_iD_i\right)}$$

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

[6] 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.