Gov 2002: 12. Causal Mechanisms

Matthew Blackwell

December 19, 2015

1. Causal Mechanisms
2. Estimands
3. Identification
4. Linear Structural Equation Models
5. Nonparametric Estimation
6. Controlled Direct Effects

1/ Causal Mechanisms

Theory and causality

- Theory $\implies$ (or $\equiv$) causal effects
- But they also tell us how those causes should impact the outcomes.
  - Theory A: causal effect is "due to" path A
  - Theory B: causal effect is "due to" path B
- How do we adjudicate between these theories when they predict the same overall effect?
- Put differently: what is the mechanism that drives a particular causal effect?
  - How do we get from cause to effect?

Example

- An example from my work is on the effect of slavery in the US South on white attitudes today.
- Whites living in former slave areas in the South today more likely to be conservative on racial issues.
- Is this due to the historical persistence of attitudes?
- Or is this effect due to demographic persistence? (More African Americans in former slave areas today $\rightsquigarrow$ whites threatened today)
- Sorting out the difference between these mechanisms is very important for our theories about political development.

What is a causal mechanisms?

- A massive diversity of definitions
- But basically: how a treatment affects an outcome
- Cannot estimate a mechanism, only test for observable implications:
  - causal mediation (effect decomposition)
  - effects modification (null effect among a subgroup)
  - presence or absence of direct effects
  - placebo tests
- Imai et al focus on the first of these, which is where our focus will be today

Notation

- Treatment variable D_i
- Outcome variable Y_i
- An intermediate, post-treatment variable, M_i, which we call a mediator.

$$\begin{array}{ccc} & M_i & \\ \nearrow & & \searrow \\ D_i & \rightarrow & Y_i \end{array}$$

Moderators vs. mediators

- Moderator: pretreatment variable that is correlated with the treatment effect.

$$\text{Cov}(\tau_i, X_i) \neq 0$$

- Mediator: a posttreatment variable that changes the effect of treatment.

Potential outcomes

- Mediators have potential outcomes $M_i(d)$: the value that the mediator takes when the treatment is d.
- Potential outcomes $Y_i(d, m)$: the value that the outcome takes when the treatment has value d and the mediator takes the value m.
- Consistency assumption to connect the potential outcomes to the observed outcomes:

$$M_i = M_i(D_i)$$
$$Y_i = Y_i(D_i, M_i(D_i))$$

Potential outcomes example

- D_i is exercise, M_i is diet, and Y_i is weight.
- d is "run 10 km/day" and m is "eat 1500 kcals"
- $Y_i(d, m)$ is the weight you would have if we forced you to run 10 km/day and eat 1500 kcals a day.

2/ Estimands

Total causal effects

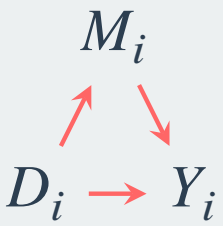
- We can recover "original" potential outcomes:

$$Y_i(d) = Y_i(d, M_i(d))$$

- Your weight if we force you to run 10 km/day, but don't intervene on your diet.
- We can define the typical individual causal effect, here called the total causal effect:

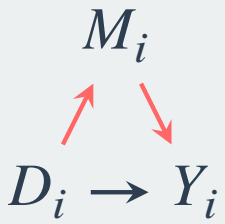
$$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

- The total causal effect allows the effect of the treatment "propogate" through all causal pathways.

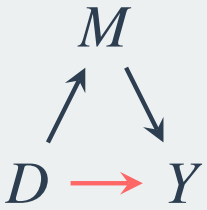


Direct and indirect effects

- The indirect effect is the part of the effect of treatment that "flows through" the mediator



- The direct effect is the part of the effect that does not flow through the mediator.



- These are loose definitions, let's be precise.

Indirect effects

- One estimand is the so-called "natural" indirect effect (NIE):

$$\delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$

  - Fix treatment to d.
  - Vary M_i by the value that it would take under treatment and control for unit i.

- If D_i doesn't affect M_i, so that $M_i(1) = M_i(0)$, then $\delta_i = 0$.
- FPOCI $\rightsquigarrow$ focus on the average natural indirect effect (ANIE):

$$\bar{\delta}(d) = \mathbb{E}[\delta_i(d)] = \mathbb{E}[Y_i(d, M_i(1)) - Y_i(d, M_i(0))]$$

Impossible counterfactuals

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0))$$

- Compare your weight when we force you to run 10 km/day and to your weight when you run 10 km/day, but keep your diet as if you didn't run at all.
- The second part, $Y_i(1, M_i(0))$, is logically unobservable.
  - Need to see you in two states of the world simultaneously, running and not running.
  - Not just the FPOCI.
  - Crossover experimental designs require strong no carry-over assumptions.
- Leads some to dismiss mediation altogether.

Natural Direct Effects

- We can also define the natural direct effect (NDE) of the treatment:

$$\eta_i(d) = Y_i(1, M_i(d)) - Y_i(0, M_i(d))$$

- Thus, the natural direct effect is the effect of moving from control to treatment while holding the mediator fixed at the value it would have under treatment status d.

When are NDEs useful?

- The canonical example: D_i is smoking, M_i is tar intake, and Y_i is lung cancer.
- We know that smoking increases tar consumption, $M_i(1) - M_i(0) > 0$,
- Also, smoking overall increases the likelihood of lung cancer, $Y_i(1, M_i(1)) - Y_i(0, M_i(0))$.
- But what would happen if we created a tar-less cigarette?
  - So that $M_i(1) = M_i(0)$ for all i.
- NDE answers this question.

Effect decomposition

- The total causal effect and the natural indirect and direct causal effects are related:

$$\tau_i = \delta_i(d) + \eta_i(1-d) = NIE_i(d) + NDE_i(1-d)$$

- Thus, we know that the ATE, $\tau = \mathbb{E}[\tau_i]$, must be the sum of the average indirect and direct effects:

$$\tau = \bar{\delta}(d) + \bar{\eta}(1-d) = ANIE(d) + ANDE(1-d)$$

- The fact that we can decompose the total effect of treatment into the sum of a direct and indirect effect if very important to social science researchers.

Other direct effects

- Another definition of direct effects is the controlled direct effect (CDE):

$$Y_i(1, m) - Y_i(0, m)$$

- The effect of running 10 km/day if we fixed your diet to 1500 kcals/day.
- ACDE is the average of these over the i units.
- In general, this effect will be different than the NDE.
  - ACDE: set M_i to m for all units
  - ANDE: set M_i to $M_i(0)$ for all units
- ACDE is identified under weaker conditions than the ANDE.

3/ Identification

Identifying indirect and direct effects

- What assumptions can identify the ANDE and ANIE?
- Imai et al use a sequential ignorability (SI) assumption, which has two parts.
  - Similar to earlier assumptions from Pearl.
  - Confusingly different from other uses of sequential ignorability by Robins and others.
- SI part 1: the treatment is independent of the potential outcomes and potential mediators, conditional on a set of covariates:

$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i | X_i = x$$

- Could be satisfied with a randomly assigned treatment

Identifying indirect and direct effects

- SI part 2: the mediator is ignorable with respect to the outcome, conditional on the treatment:

$$Y_i(d', m) \perp\!\!\!\perp M_i(d) | D_i = d, X_i = x$$

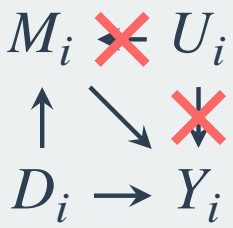
- This must hold for all values of d, d'.
- Note that we have to believe ignorability in certain cross-world comparisons:

$$Y_i(1, m) \perp\!\!\!\perp M_i(0) | D_i = 0, X_i = x$$

- Could be satisfied by randomizing M_i, but then the effect of D_i is not "natural."

SI and posttreatment bias

- SI assumes that posttreatment bias is not a problem.
- The mediator is as-if random, so these situations can never happened:

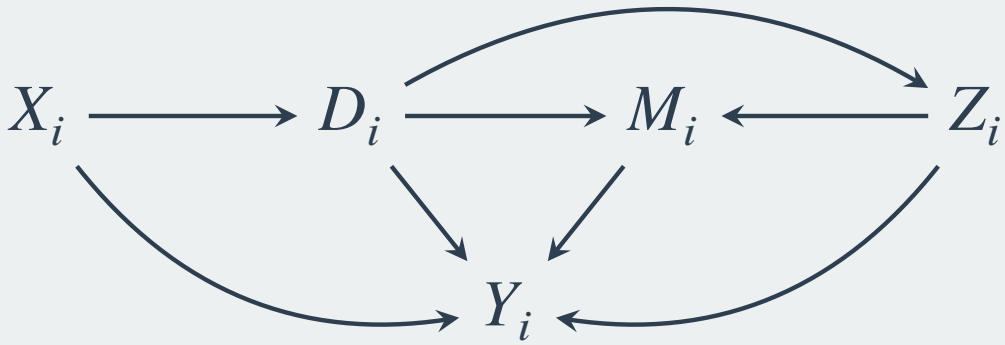


- Never any collider bias.
- Is this plausible? It depends on the application.

Limitations of sequential ignorability

$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i | X_i = x$$
$$Y_i(d', m) \perp\!\!\!\perp M_i(d) | D_i = d, X_i = x$$

- Conditioning set X_i is the same for both stages.
- What if there are confounders for the relationship between M and Y that are affected by D? Too bad!



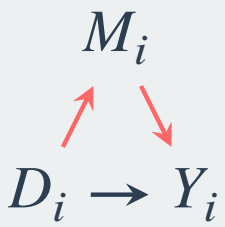
- More on this in a bit.

Identifying (in)direct effects

- Under SI and consistency, we can write the ANIE as a function of the observed data.
- With a binary mediator and a binary treatment:

$$\begin{aligned}\bar{\delta}(d) =& \{\mathbb{P}[M_i = 1|D_i = 1, X_i] - \mathbb{P}[M_i = 1|D_i = 0, X_i]\} \\ & \cdot \{\mathbb{E}[Y_i|M_i = 1, D_i = d, X_i] - \mathbb{E}[Y_i|M_i = 0, D_i = d, X_i] \\ =& (\text{effect of } D_i \text{ on } M_i) \times (\text{effect of } M_i \text{ on } Y_i)\end{aligned}$$

- Intuitive given the DAG:



(In)direct effects with non-binary mediators

- Let's say that the mediator has J categories:

$$ANIE(d) = \sum_{m=0}^{J-1} \mathbb{E}[Y_i|M_i = m, D_i = d, X_i] \cdot \{\mathbb{P}[M_i = m|D_i = 1, X_i] - \mathbb{P}[M_i = m|D_i = 0, X_i]\}$$

- The ANDE is the following:

$$ANDE(d) = \sum_{m=0}^{J-1} (\mathbb{E}[Y_i|M_i = m, D_i = 1, X_i] - \mathbb{E}[Y_i|M_i = m, D_i = 0, X_i]) \cdot \{\mathbb{P}[M_i = m|D_i = d, X_i]\}$$

- The ANDE is the effect of D_i for a fixed m, averaged over the distribution of M_i when $D_i = 0$.

Alternative identification

- Robins proposed a different identification strategy, based on a no-interactions assumption:

$$Y_i(1, m) - Y_i(0, m) = Y_i(1, m') - Y_i(0, m')$$

- The CDE does not depend on m for any unit i.
- $\rightsquigarrow$ ACDE $=$ ANDE.
- Strong assumption because it has to hold at the individual level (like monotonicity for IV).

4/ Linear Structural Equation Models

Estimation

- Let's say that we have a linear, structural model for all variables:

$$M_i(d) = \alpha_0 + \alpha_1 d + \eta_i$$
$$Y_i(d, m) = \beta_0 + \beta_1 d + \beta_2 m + \varepsilon_i$$

- Here the effect of treatment and mediator are constant across units.
- This is a huge simplification and may be incorrect.
- Allows us to "plug-in" and get potential outcomes:

$$\begin{aligned} Y_i(1, M_i(1)) &= \beta_0 + \beta_1 \times 1 + \beta_2 M_i(1) + \varepsilon_i \\ &= \beta_0 + \beta_1 \times 1 + \beta_2 \left(\alpha_0 + \alpha_1 \times 1 + \eta_i\right) + \varepsilon_i \end{aligned}$$

Linear models and mediation

- It's clear that we can write the total effect of the treatment in the following way:

$$\begin{aligned} Y_i(1, M_i(1)) - Y_i(0, M_i(0)) =& \beta_0 + \beta_1 + \beta_2(\alpha_0 + \alpha_1 + \eta_i) + \varepsilon_i \\ & - \beta_0 - \beta_2(\alpha_0 + \eta_i) - \varepsilon_i \\ =& \beta_1 + \beta_2 \cdot \alpha_1 \end{aligned}$$

- What about the indirect effect:

$$\begin{aligned} Y_i(0, M_i(1)) - Y_i(0, M_i(0)) =& \beta_0 + \beta_2(\alpha_0 + \alpha_1 + \eta_i) + \varepsilon_i \\ & - \beta_0 - \beta_2(\alpha_0 + \eta_i) - \varepsilon_i \\ =& \beta_2 \cdot \alpha_1 \end{aligned}$$

Estimation with LSEMs

- Estimate the total effect from a regression of Y_i on D_i and X_i
- Estimate the $\widehat{\beta}_1$ and $\widehat{\beta}_2$ from a regression of Y_i on D_i, M_i, and X_i.
- Estimate $\widehat{\alpha}_1$ from a regression of M_i on D_i
- Direct effect is $\widehat{ANDE} = \widehat{\beta}_1$
- Indirect effect as the product: $\widehat{ANIE} = \widehat{\alpha}_1 \widehat{\beta}_2$.

Interactions

- Implicit assumption: no interactions

$$ANIE(1) = ANIE(0)$$

- We could incorporate an interaction into the model here to allow for the indirect effect to vary.

$$Y_i(d, m) = \beta_0 + \beta_1 d + \beta_2 m + \beta_3 dm + \varepsilon_i$$

Variance estimates

- The variance of the total effect and the direct effect are straightforward.
  - Just the SE of the estimated coefficients.
- The indirect effect is more complicated because it is a function of multiple parameters.
- Using the delta method, the variance of $\widehat{ANIE} = \widehat{\alpha}_1 \widehat{\beta}_2$ can be written:

$$\mathbb{V}[\widehat{ANIE}] \approx \widehat{\alpha}_1^2 \mathbb{V}[\widehat{\beta}_2] + \widehat{\beta}_2^2 \mathbb{V}[\widehat{\alpha}_1]$$

- We can use this formula to estimate standard errors for the indirect effects.

5/ Nonparametric Estimation

Nonparametric estimation

- LSEMs require strong modeling assumptions $\rightsquigarrow$ what about nonparametrics?
- If the number of categories in M_i, D_i, and X_i are small, use plug-in estimator for the CEF of Y_i:

$$\widehat{\mathbb{E}}[Y_i|M_i = m, D_i = d, X_i = x] = \frac{\sum_i Y_i \mathbb{1}\{M_i = m, D_i = d, X_i = x\}}{\sum_i \mathbb{1}\{M_i = m, D_i = d, X_i = x\}}$$

- Same for M_i:

$$\widehat{\mathbb{P}}[M_i = m|D_i = d, X_i = x] = \frac{\sum_i \mathbb{1}\{M_i = m, D_i = d, X_i = x\}}{\sum_i \mathbb{1}\{D_i = d, X_i = x\}}$$

What about more complicated scenarios?

- If the number of categories is large, then we can use nonparametric regressions for the outcome and the mediator.

$$\mu_{dm}(x) = \mathbb{E}[Y_i | M_i = m, D_i = d, X_i = x]$$

- Flexibly estimate $\mu_{dm}(x)$ as a function of x using splines of x.
- To get the standard errors, we can use bootstrapping.
- Need to be careful with the curse of dimensionality in X_i. Use good nonparametric strategies (cross-validation, etc)

Continuous mediator, nonparametric

- What if the mediator is continuous? Things get tricky.
- Need to integrate over the distribution of the mediators to get the ANIE:

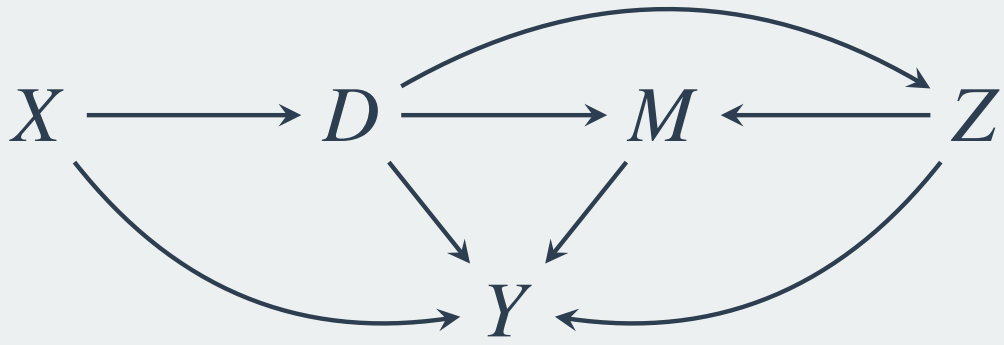
$$\bar{\delta}(d) = \int \int \mathbb{E}[Y_i | M_i = m, D_i = d, X_i = x] \{dF_{M_i|D_i=1,X_i=x}(m) - dF_{M_i|D_i=0,X_i=x}(m)\} dF_{X_i}(x)$$

- Obviously, this is a much harder problem. In this case, we actually can use Monte Carlo simulation to take the integral.
- Modeling M_i probably appropriate here.

6/ Controlled Direct Effects

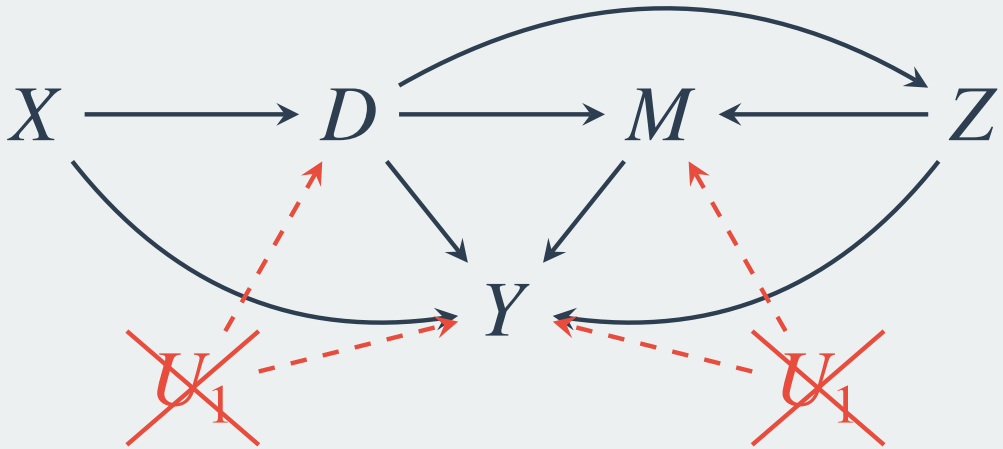
Intermediate confounders

- Intermediate confounders are variables that confound the $M_i \to Y_i$ relationship, but are affected by D_i
- Here we represent them as Z_i:



- Can also be thought of as other mediators, about which we aren't directly interested.
- Avin, Shpitser and Pearl (2003) showed that ANDE/ANIE identification not possible when SI incorporates intermediate confounders.

Sequential ignorability, II



- New version of sequential ignorability with intermediate confounders:

$$Y_i(d,m) \perp\!\!\!\perp D_i | X_i$$
$$Y_i(d,m) \perp\!\!\!\perp M_i | Z_i, D_i, X_i$$

- No unmeasured confounders for D_i conditional on X_i
- No unmeasured confounders for M_i conditional on Z_i, D_i, X_i

Sequential ignoribablity, II

$$Y_i(d, m) \perp\!\!\!\perp D_i | X_i$$
$$Y_i(d, m) \perp\!\!\!\perp M_i | Z_i, D_i, X_i$$

- Original Robins definition of sequential ignorability.
- No cross-world assumptions, allows for intermediate confounders.
- Will only allow for the identification of the ACDE:

$$ACDE(m) = \mathbb{E}[Y_i(1, m) - Y_i(0, m)]$$

- Require Robins's no-interaction assumption to connect ACDE to ANDE.

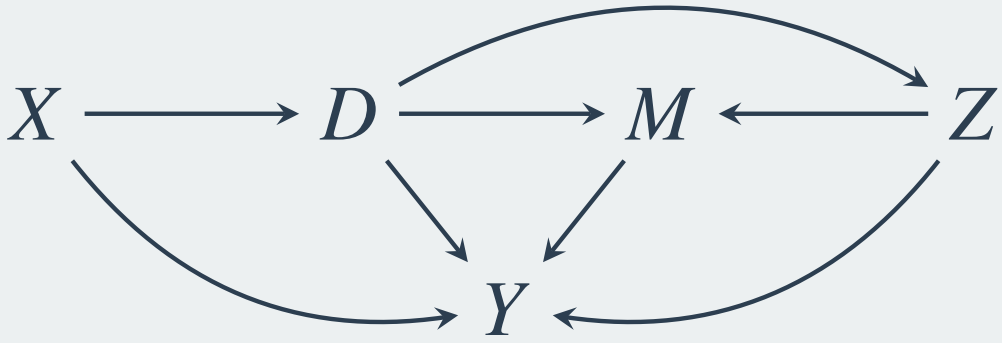
Identifying the ACDE

- Nonparametric idenfication of the ACDE:

$$
\begin{aligned}
&\mathbb{E}[Y_i(d,m)] \\
&\quad = \int_x \mathbb{E}[Y_i(d,m)|x]dF_X(x) \quad \text{(LIE)} \\
&\quad = \int_x \mathbb{E}[Y_i(d,m)|x,d]dF_X(x) \quad \text{(n.u.c for D)} \\
&\quad = \int_x \int_z \mathbb{E}[Y_i(d,m)|x,d,z]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(LIE)} \\
&\quad = \int_x \int_z \mathbb{E}[Y_i(d,m)|x,d,z,m]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(n.u.c for M)} \\
&\quad = \int_x \int_z \mathbb{E}[Y_i|x,d,z,m]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(consistency)}
\end{aligned}
$$

- Everything in the last line is identified from the data.
- Relationship can generalized to any number of treatments, and is called the g-formula by Robins.

Estimating direct effects



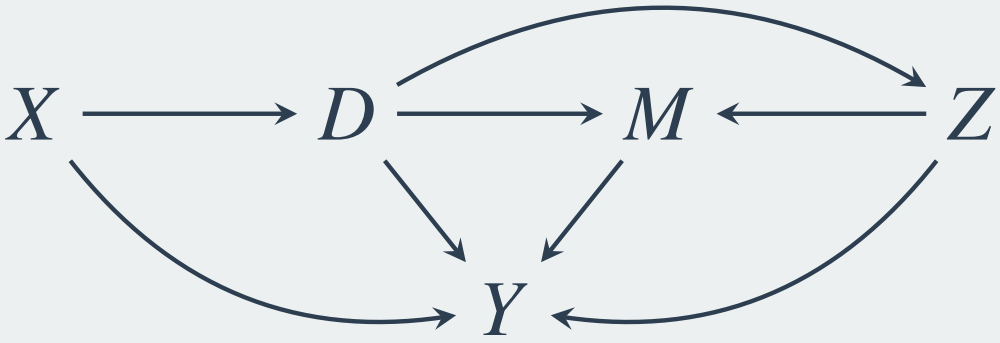
- Controlling for Z_i and $M_i \rightsquigarrow$ posttreatment bias
  - Conditioning on a collider $\rightsquigarrow$ selection bias
  - Conditioning on $Z_i \rightsquigarrow$ masking part of the CDE
- Compare this conditioning approach:

$$\mathbb{E}[Y_i|x, d = 1, z, m] - E[Y_i|x, d = 0, z, m]$$

- And the identification result from the g-formula:

$$\int_x \int_z \mathbb{E}[Y_i|x, d = 1, z, m] dF_{Z|D,X}(z|d = 1, x) dF_X(x)$$
$$- \int_x \int_z \mathbb{E}[Y_i|x, d = 0, z, m] dF_{Z|D,X}(z|d = 0, x) dF_X(x)$$

Sequential g-estimation

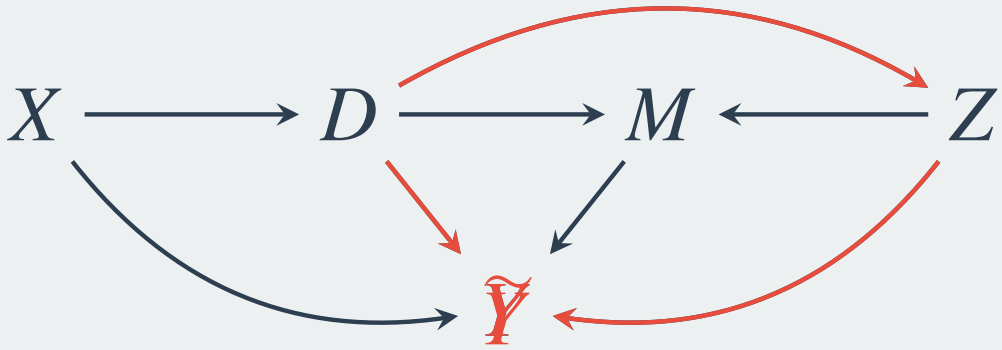


- Sequential g-estimation is one of many approaches in these settings.
  - Other approaches include weighting.
- Run the "long" regression:

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i' \gamma_3 + Z_i' \gamma_4 + \varepsilon_i$$

- γ_1 is not the CDE (posttreatment bias)
- γ_2 is the effect of M_i on Y_i

Blip down



- Create a blipped down (or demediated) outcome: $\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$
- The blip-down removes the effect of M_i on Y_i from the outcome.
- Any remaining effect of D_i on Y_i is just the CDE:

$$\mathbb{E}[\widetilde{Y}_i | D_i = d, X_i] = \mathbb{E}[Y_i(d, 0) | X_i]$$

Sequential g-estimation

1. Run a regression of Y_i on M_i, Z_i, D_i, X_i.

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i' \gamma_3 + Z_i' \gamma_4 + \varepsilon_i$$

2. Subtract off the effect of M_i on Y_i:

$$\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$$

3. Regress blipped-down outcome on D_i and X_i:

$$\widetilde{Y}_i = \beta_0 + \beta_1 D_i + X_i' \beta_2 + \eta_i$$
$$CDE(0) = \mathbb{E}[Y_i(1,0) - Y_i(0,0)] = \beta_1$$

4. Bootstrap or complicated variance estimator for SEs
   - Second regression ignores the first regression.

Notes on sequential g-estimation

- Relies on a no (average) interaction assumption between CDE and intermediate confounders.
- We can weaken this, but requires us to model the distribution of Z_i which might be very high dimensional:

$$\int_x \int_z \mathbb{E}[Y_i|x, d = 1, z, m] dF_{Z|D,X}(z|d = 1, x) dF_X(x)$$
$$- \int_x \int_z \mathbb{E}[Y_i|x, d = 0, z, m] dF_{Z|D,X}(z|d = 0, x) dF_X(x)$$

- Typical selection on observables: need correct model for covariates in both steps.
- ATE - ACDE $\neq$ an indirect effect, but still can tell us something about mechanisms.

Wrap-up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.
    - Mediation
    - Controlled direct effects
    - Effect modification
    - Placebo tests