14. Causal Mechanisms

ISS5096 || ECI
Jaewon (“Jay-one”) Yoo
National Tsing Hua University

Roadmap

1. Causal Mechanisms

2. Estimands

3. Identification

4. Estimation
   - Linear Structural Equation Models
   - Nonparametric Estimation
   - Estimating Controlled Direct Effects

5. Mechanisms in Practice

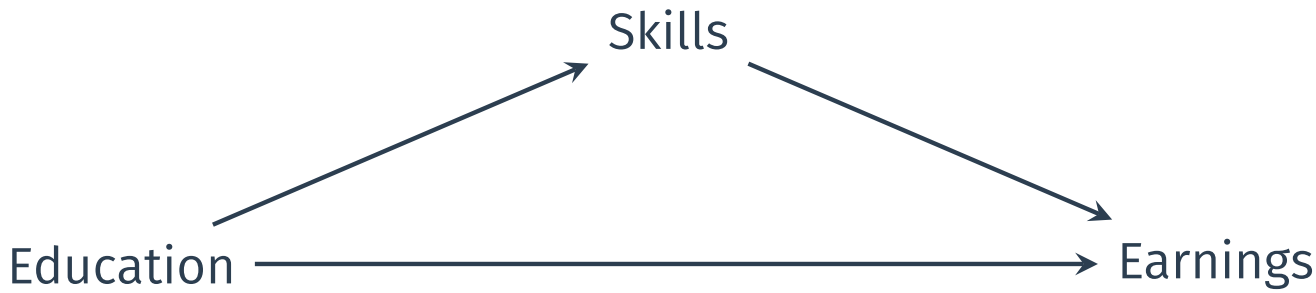
6. Wrap-Up

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 to adjudicate between theories that predict the same ATE?
- Put differently: what **mechanism** drives a particular causal effect?

Example: Why Does Education Raise Earnings?



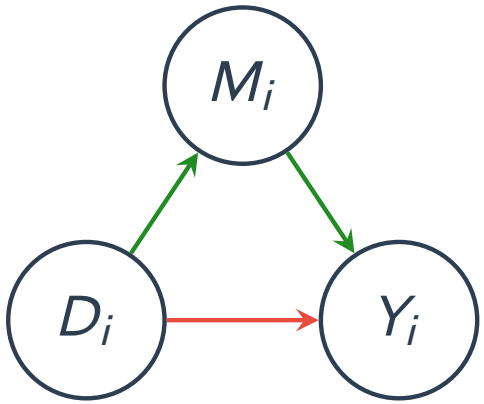
- The education earnings premium: among the best-documented effects in social science. But *why* does education pay?
- Two classic theories, same total effect:
  - **Human capital** (Becker 1964): education builds productive skills ⇝ operates *through* skill formation.
  - **Signaling** (Spence 1973): a degree certifies preexisting ability to employers ⇝ operates *not through* skills.
- Same ATE, opposite policy implications: subsidize skill formation vs. rethink credential requirements.

What is a Causal Mechanism?

- 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

Notation

- DAG representation:
  - Treatment variable $D_i \in \{0, 1\}$
  - Mediator, $M_i \in \mathcal{M}$
  - Potential outcome variable $Y_i(d, m)$



- Mediation goal: decompose total effect into direct and indirect effects.
- Moderator vs mediator:
  - **Moderator**: pretreatment variable correlated with the treatment effect.
  - **Mediator**: a posttreatment variable that changes the treatment effect.
  - Mediator has potential outcomes as well: $M_i(d)$
- Consistency: $M_i = M_i(D_i)$ and $Y_i(D_i, M_i(D_i))$.

Interpreting the Potential Outcomes

- Example: D_i is exercise, M_i is diet, and Y_i is weight.
  - $D_i = 1$ is "run 10 km/day", $D_i = 0$ is don't run
  - M_i is the number of calories to eat.

- Some different possible potential outcomes:
  - $Y_i(1, 1500)$: weight you would have if we forced you to run 10 km/day and eat 1500 kcals a day.
  - $Y_i(1, M_i(1))$: weight if you run 10 km/day, but no intervention on diet.
  - $Y_i(0, M_i(1))$: weight if you didn't run, but ate like you did.

- Cross-world counterfactuals $Y_i(0, M_i(1))$ logically impossible to observe.
  - Not just the fundamental problem of CI.

2/ Estimands

Controlled Direct Effects (CDE)

- Definition for each $m \in \mathcal{M}$:

$$\text{Individual:} \quad \xi_i(m) = Y_i(1, m) - Y_i(0, m)$$
$$\text{Average:} \quad \overline{\xi}(m) = \mathbb{E}\left[Y_i(1, m) - Y_i(0, m)\right]$$

- Interpretation:
  - Effect of treatment when holding mediator fixed at m.
  - The effect of running 10 km/day if we fixed your diet to 1500 kcals/day.
  - Target of experiment manipulating D_i and M_i.

- If M_i fully mediates effect of D, then CDEs will be 0 for all m.
  - $\rightsquigarrow$ can be used to establish existence of unmediated path from $D \longrightarrow Y$.

- Can capture **interactions** if $\xi_i(m) \neq \xi_i(m')$

Natural Indirect Effects (NIE)

- Definition of the **natural indirect effect** (NIE):

$$\text{Individual:} \quad \delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$
$$\text{Average:} \quad \overline{\delta}(d) = \mathbb{E}\left[Y_i(d, M_i(1)) - Y_i(d, M_i(0))\right]$$

- Interpretation:
  - Effect of the a change in the mediator induced by the effect of D_i on M_i.
  - Holding fixed the value of treatment.

- Also called the **causal mediation effect**

- If D_i doesn't affect M_i, so that $M_i(1) = M_i(0)$, then $\delta_i = 0$.

Natural Direct Effects (NDEs)

- Definition of the **natural direct effect** (NDE) of the treatment:

$$\text{Individual:} \quad \zeta_i(d) = Y_i(1, M_i(d)) - Y_i(0, M_i(d))$$
$$\text{Average:} \quad \overline{\zeta}(d) = \mathbb{E}\left[Y_i(1, M_i(d)) - Y_i(0, M_i(d))\right]$$

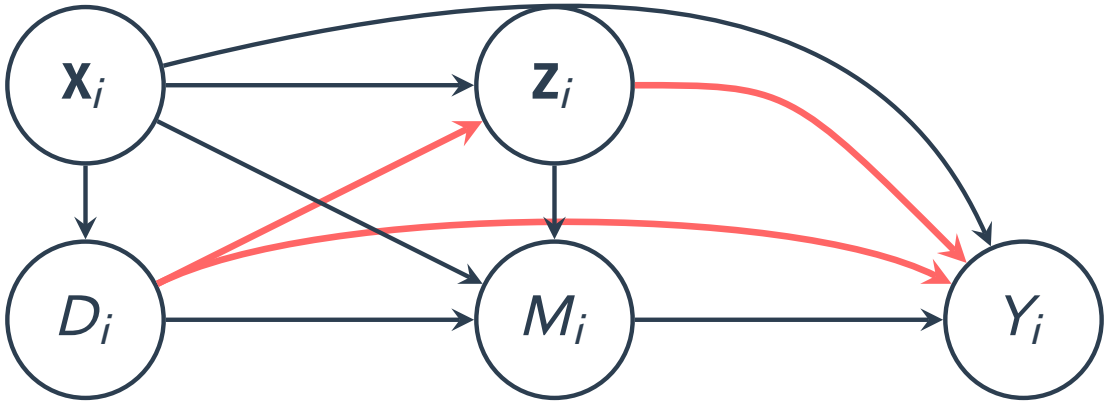
- Interpretation:
  - Effect of treatment when mediator is at its natural value for $D_i = d$.
  - Effect of a redesigned treatment that doesn't affect the mediator

- Total effect decomposition:

$$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \underbrace{\delta_i(d)}_{\text{NIE}} + \underbrace{\zeta_i(1-d)}_{\text{NDE}}$$

3/ Identification

Identification for CDEs



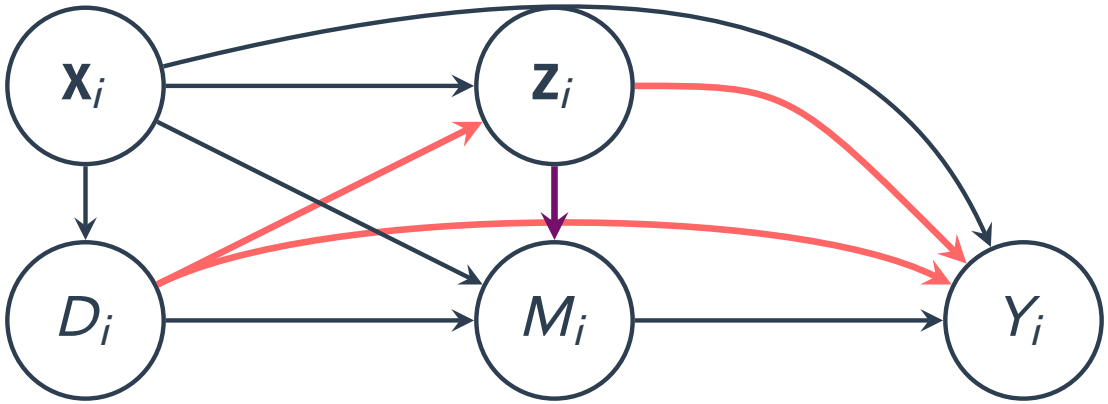
- Conditioning sets:
  - $\mathbf{X}_i$: pre-treatment confounders
  - $\mathbf{Z}_i$: post-treatment or intermediate confounders
- **Sequential ignorability** (Robins):

$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$

$$Y_i(d, m) \perp\!\!\!\perp M_i \mid \mathbf{X}_i = \mathbf{x}, D_i = d, \mathbf{Z}_i = \mathbf{z}$$

- Interpretation: two "selection-on-observables" assumptions.
  - D_i randomly assigned conditional on $\mathbf{X}_i$.
  - M_i randomly assigned conditional on $\mathbf{X}_i$, D_i, and $\mathbf{Z}_i$.

The Dual Role of $\mathbf{Z}_i$



- $\mathbf{X}_i$ is **pre-treatment**: all of its roles point the same way $\rightsquigarrow$ adjust (selection on observables, Week 5).
- $\mathbf{Z}_i$ is **post-treatment**: its two roles **conflict**.
  - **Confounder of** $M_i \to Y_i$: given $(\mathbf{X}_i, D_i)$, the back door $M_i \leftarrow \mathbf{Z}_i \to Y_i$ stays open $\rightsquigarrow$ must *adjust* for $\mathbf{Z}_i$.
  - **Outcome of** D_i: the path $D_i \to \mathbf{Z}_i \to Y_i$ is *part of the CDE* $\rightsquigarrow$ must *not block* it.
- No single conditioning set can do both $\rightsquigarrow$ condition on $\mathbf{Z}_i$ *inside the conditional mean*, then average it back out at its treatment-specific distribution (next slide).

Identifying the ACDE

- Post-treatment bias if we just condition on $\mathbf{Z}_i$:

$$\overline{\xi}(m) \neq \sum_{\mathbf{x},\mathbf{z}} \{\mathbb{E}[Y_i \mid D_i = 1, M_i = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}] - \mathbb{E}[Y_i \mid D_i = 0, M_i = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}]\}\mathbb{P}(\mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z})$$

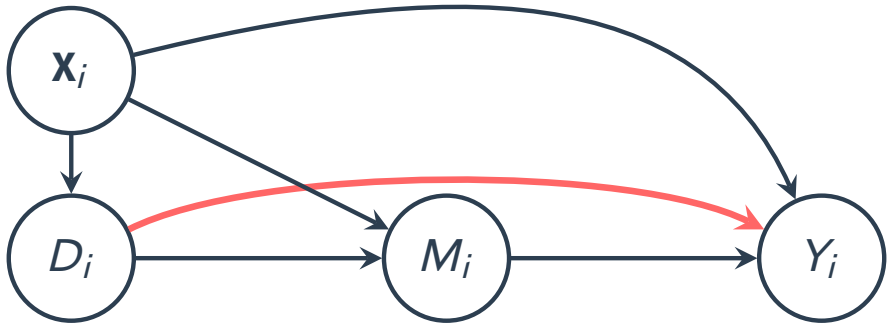
  - Ignores that $\mathbf{Z}_i$ depends on D_i!

- Nonparametric identification of the ACDE:

$$\overline{\xi}(m) = \sum_{\mathbf{x},\mathbf{z}} \{\mathbb{E}[Y_i \mid D_i = 1, M_i = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}]\mathbb{P}(\mathbf{Z}_i = \mathbf{z} \mid D_i = 1, \mathbf{X}_i = \mathbf{x}) - \mathbb{E}[Y_i \mid D_i = 0, M_i = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}]\mathbb{P}(\mathbf{Z}_i = \mathbf{z} \mid D_i = 0, \mathbf{X}_i = \mathbf{x})\} \times \mathbb{P}(\mathbf{X}_i = \mathbf{x})$$

- **g-formula** (Robins) generalizes to any number of treatments

Identification for Mediation



- **Sequential ignorability** (Imai et al. 2010):

$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$
$$Y_i(d, m) \perp\!\!\!\perp M_i \mid \mathbf{X}_i = \mathbf{x}, D_i = d$$

  - $\rightsquigarrow$ ANIE and ANDE are identified.
- **No post-treatment confounders** (measured or unmeasured)
  - Why? The cross-world conditional $\mathbb{E}[Y_i(1, m) \mid M_i(0) = m]$ is confounded by any $\mathbf{Z}_i$:

$$M_i(0) \longleftarrow \mathbf{Z}_i(0) \underset{\text{same unit}}{\longleftrightarrow} \mathbf{Z}_i(1) \longrightarrow Y_i(1, m)$$

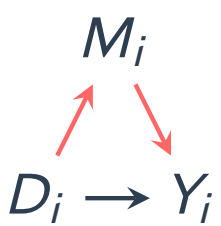
  - The middle link joins two versions of *one* variable $\rightsquigarrow$ nothing observable blocks it: a **recanting witness** (Avin, Shpitser, and Pearl 2005), not identified, measured or not.

Identifying (In)direct Effects

- ANIE under binary treatment/mediator:

$$\bar{\delta}(d) = \sum_{\mathbf{x}} \Big(\{ \underbrace{\mathbb{P}[M_i = 1 \mid D_i = 1, \mathbf{X}_i = \mathbf{x}] - \mathbb{P}[M_i = 1 \mid D_i = 0, \mathbf{X}_i = \mathbf{x}]}_{\text{effect of } D_i \text{ on } M_i} \}$$
$$\times \{ \underbrace{\mathbb{E}[Y_i \mid M_i = 1, D_i = d, \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i \mid M_i = 0, D_i = d, \mathbf{X}_i = \mathbf{x}]}_{\text{effect of } M_i \text{ on } Y_i} \} \Big)$$
$$\times \mathbb{P}(\mathbf{X}_i = \mathbf{x})$$

- Multiply paths given $\mathbf{X}_i$ and aggregate intuitive given DAG:



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/ Estimation

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

Interactions

- **Implicit assumption**: no interactions

$$\text{ANIE}(1) = \text{ANIE}(0)$$

- 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 $\widehat{\text{ANIE}} = \widehat{\alpha}_1 \widehat{\beta}_2$ can be written:

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

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

Nonparametric Estimation

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

$$\widehat{\mathbb{E}}[Y_i \mid M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}] = \frac{\sum_i Y_i \mathbb{1}\{M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}\}}{\sum_i \mathbb{1}\{M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}\}}$$

- Same for M_i:

$$\widehat{\mathbb{P}}[M_i = m \mid D_i = d, \mathbf{X}_i = \mathbf{x}] = \frac{\sum_i \mathbb{1}\{M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}\}}{\sum_i \mathbb{1}\{D_i = d, \mathbf{X}_i = \mathbf{x}\}}$$

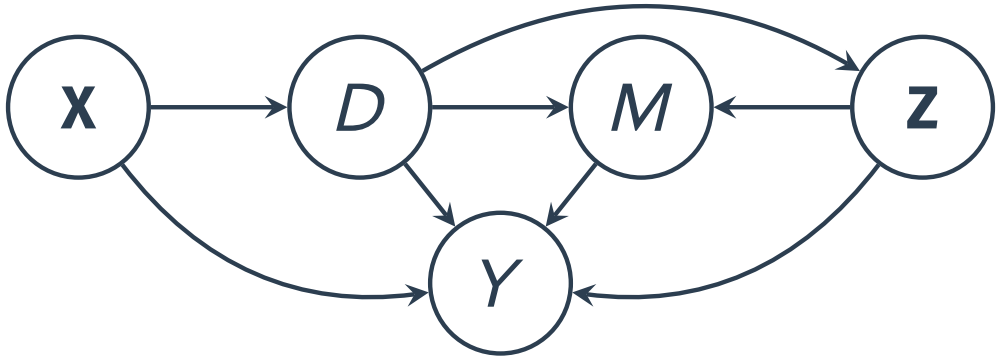
What About More Complicated Scenarios?

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

$$\mu_{dm}(\mathbf{x}) = \mathbb{E}[Y_i \mid M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}]$$

- Flexibly estimate $\mu_{dm}(\mathbf{x})$ as a function of $\mathbf{x}$ using splines of $\mathbf{x}$.
  - Spline: split the range of $\mathbf{x}$ at **knots**, fit a low-degree polynomial within each segment, join smoothly $\rightsquigarrow$ flexible like a high-degree polynomial, without its wild global behavior.
- To get the standard errors, we can use bootstrapping.
- Need to be careful with the curse of dimensionality in $\mathbf{X}_i$. Use good nonparametric strategies (cross-validation, etc)

Sequential G-Estimation

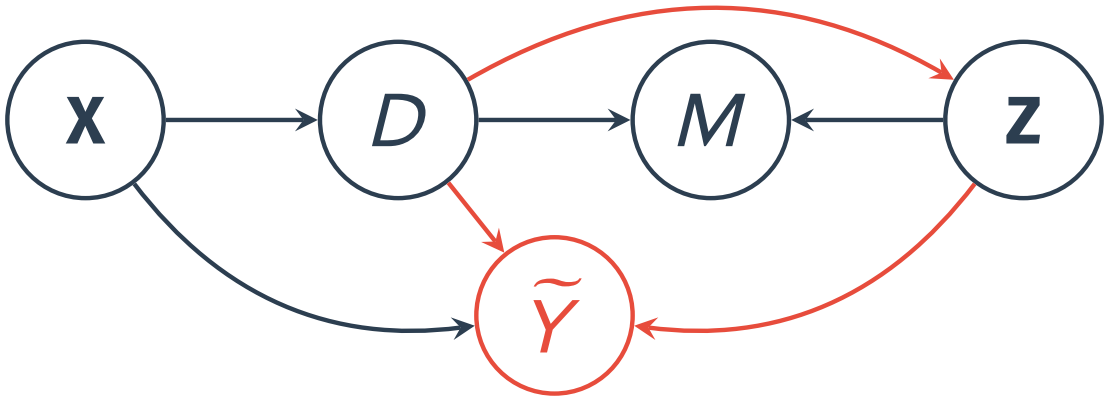


- Back to the ACDE: the **g-formula** identified it; now we estimate it. **Sequential g-estimation** recovers the g-formula's answer with two regressions, instead of modeling the distribution of $\mathbf{Z}_i$.
  - One of Robins' **g-methods**; other routes: computing the integral directly, weighting.
  - Linear version of a broader class called **structural nested mean models**
- Run the "long" regression:

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + \mathbf{X}_i'\gamma_3 + \mathbf{Z}_i'\gamma_4 + \varepsilon_i$$

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

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:

$$\mathbb{E}[\widetilde{Y}_i \mid D_i = d, \mathbf{X}_i] = \mathbb{E}[Y_i(d, 0) \mid \mathbf{X}_i]$$

  - Relies on correct modeling of the outcome!

Sequential G-Estimation

1. Run a regression of Y_i on $M_i, \mathbf{Z}_i, D_i, \mathbf{X}_i$.

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + \mathbf{X}_i'\gamma_3 + \mathbf{Z}_i'\gamma_4 + \varepsilon_i$$

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 $\mathbf{X}_i$:

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

4. Bootstrap, or the consistent two-step variance estimator derived in Acharya, Blackwell, and Sen (2016)
    - Second regression ignores the first regression.

5/ Mechanisms in Practice

Two Questions, Two Packages

- Everything so far has an off-the-shelf R implementation. Two workflows, matching the two estimands:
  - **Decompose** the total effect into (in)direct parts (NIE/NDE) $\rightsquigarrow$ `mediation` (Imai, Keele, Tingley & Yamamoto).
  - Estimate the **controlled direct effect** via sequential g $\rightsquigarrow$ `DirectEffects` (Acharya, Blackwell & Sen).
- The LSEM product $\widehat{\text{ANIE}} = \widehat{\alpha}_1 \widehat{\beta}_2$ is just the simplest case of what `mediation` automates. It adds inference and a sensitivity analysis on top.
- Mediation example, `framing` (Brader, Valentino & Suhay 2008): does a negative, Latino-cued immigration story (D) drive anti-immigration action (Y, cong_mesg) *through* anxiety (M, emo)?

Step 1: Two Models ⇝ One `mediate()` Call

```
> library(mediation)
> data("framing", package = "mediation")

> # (1) mediator model:  M ~ D + X
> med.fit <- lm(emo ~ treat + age + educ + gender + income,
                data = framing)

> # (2) outcome model:   Y ~ M + D + X   (probit, binary Y)
> out.fit <- glm(cong_mesg ~ emo + treat + age + educ + gender + income,
                 data = framing, family = binomial("probit"))

> # (3) decompose; set seed since mediate() simulates
> set.seed(2014)
> med.out <- mediate(med.fit, out.fit, treat = "treat",
                     mediator = "emo", robustSE = TRUE, sims = 1000)
```

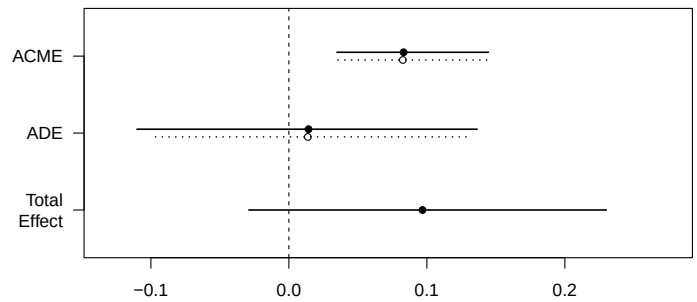
- `mediate()` **simulates** the potential outcomes from the two models; no closed-form product needed.
- Add `emo * treat` for a treatment–mediator interaction (the lecture's interaction case); we keep it additive so the closed-form `medsens` applies.

Step 2: Reading the Decomposition

```
> summary(med.out)

                        Estimate  95% Lo  95% Hi p-value
ACME (control)            0.0826  0.0356  0.1440  <2e-16 ***
ACME (treated)            0.0831  0.0348  0.1446  <2e-16 ***
ADE (control)             0.0137 -0.0967  0.1318   0.818
ADE (treated)             0.0142 -0.1101  0.1365   0.818
Total Effect              0.0968 -0.0290  0.2301   0.136
Prop. Mediated (control)  0.7706 -6.3968  4.7049   0.136
Prop. Mediated (treated)  0.7938 -5.7506  4.5154   0.136
ACME (average)            0.0829  0.0351  0.1444  <2e-16 ***
ADE (average)             0.0140 -0.1047  0.1343   0.818
Prop. Mediated (average)  0.7822 -6.0737  4.6101   0.136

Sample Size Used: 265     Simulations: 1000
```



`> plot(med.out)`

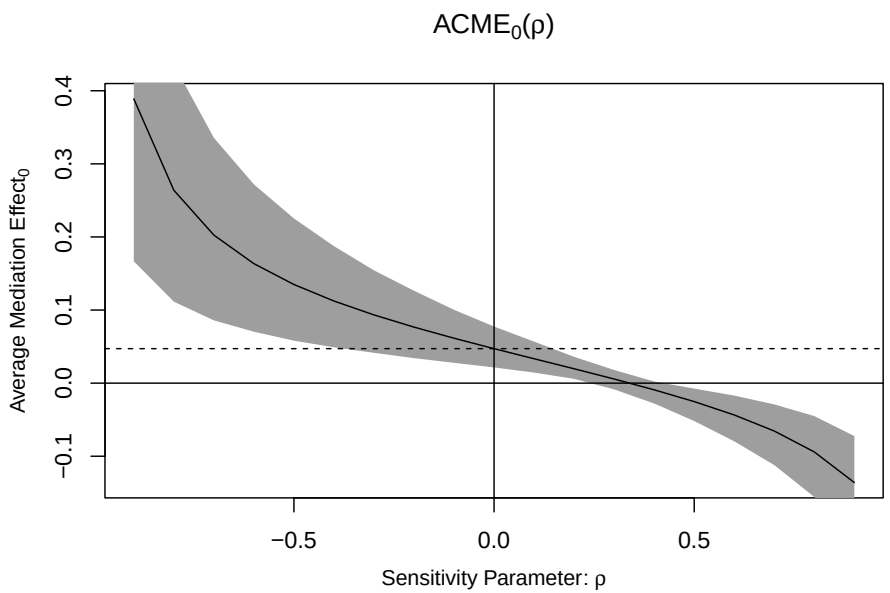
- Output ⟶ slide notation: **ACME** $= \bar{\delta}(d)$ (NIE), **ADE** $= \bar{\zeta}(d)$ (NDE); (control)/(treated) $= d = 0, 1$, unequal only through the probit's nonlinearity ⇝ read the (average) rows.
- **ACME**'s CI excludes 0; **ADE** and the *total* effect cover it ⇝ the detectable channel is the indirect one.
- `Prop. Mediated` divides by that shaky total ⇝ CI -6.07 to 4.61. Read the channels; don't headline the ratio.

Step 3: The Result Rests on an Untestable Assumption

- Identification needs **sequential ignorability**: no unmeasured confounder of the $M \longrightarrow Y$ link, given $D_i, \mathbf{X}_i$.
- Randomizing D does *not* buy this: the mediator is never randomized.
- `medsens()` asks how big a violation overturns the result, indexed by $\rho = \text{cor}(\varepsilon_M, \varepsilon_Y)$:

```
> sens.out <- medsens(med.out, rho.by = 0.1, effect.type = "indirect")
> summary(sens.out)
> plot(sens.out, sens.par = "rho")
```

Step 3: How Fragile Is the Story?



- Estimated ACME (dashed) hits **zero at** $\rho \approx 0.3$.
- Equivalently, an unobserved confounder explaining $\approx 9\%$ of the residual variation in *both* M and Y ($R^2_M \cdot R^2_Y \approx 0.09$) erases the **entire** indirect effect.
- We cannot test $\rho = 0$, so the decomposition is only as good as that assumption.

Controlled Direct Effects: `DirectEffects`

- A different question: not "how much runs *through* M?" but "is there *any* effect **not** through M?" $\rightsquigarrow$ the **ACDE**.
- Sequential g-estimation (Acharya, Blackwell & Sen 2016), exactly the recipe from the last section:
  1. Regress Y on $D, M, \mathbf{Z}, \mathbf{X}$; read off $\widehat{\gamma}$ on M.
  2. Blip down: $\widetilde{Y} = Y - \widehat{\gamma}\, M$.
  3. Regress $\widetilde{Y}$ on $D, \mathbf{X}$ $\rightsquigarrow$ the coefficient on D is the ACDE.
- Buys weaker assumptions (allows treatment-induced confounding of M); the price is **no decomposition**: no "% mediated".

Sequential-g in Code: the Plough Hypothesis

- ploughs (Alesina, Giuliano & Nunn 2013): does plough agriculture depress women's political participation *directly*, or only via income?

```
> library(DirectEffects)
> data(ploughs)

> # 3-part formula:  Y ~ baseline X + treatment | intermediate Z | mediator M
> form <- women_politics ~ plow + agricultural_suitability + tropical_climate +
            large_animals + political_hierarchies + economic_complexity + rugged |
          years_civil_conflict + years_interstate_conflict + oil_pc +
            european_descent + communist_dummy + polity2_2000 + serv_va_gdp2000 |
          centered_ln_inc + centered_ln_incsq      # mediator: (log) income

> direct <- sequential_g(form, data = ploughs)
> summary(direct)
```

The Controlled Direct Effect

```
> summary(direct)

t test of coefficients:

            Estimate Std.Err t value Pr(>|t|)
(Intercept) 12.1845   3.6444   3.343  0.0011 **
plow        -4.8388   2.3447  -2.064  0.0413 *
...
```

- **ACDE** of the plough = −4.84 ($p = 0.04$).
- A direct effect on women's participation **survives** netting out income, not *only* an economic-development story.
- Sensitivity to intermediate-confounder bias: `cdesens()`, the CDE analogue of `medsens`.

6/ Wrap-Up

Wrap-Up: Why Mechanisms Stay Contested

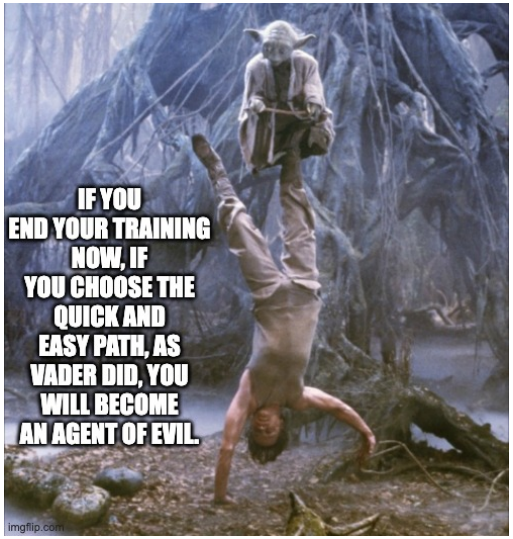
- **Mechanisms are hard**, and unusually so:
  - Mediation needs **sequential ignorability**: untestable, and *not* delivered by randomizing D.
  - Natural (in)direct effects invoke **cross-world** potential outcomes $Y_i(d, M_i(d'))$ we can never observe.
- The people who built these tools say the same. Kosuke Imai (Harvard), in his *2025 NBER Methods Lecture*, is blunt:
  - "Even when M is randomized, NIE/NDE are unidentifiable; sensitivity analysis plays an important role."
  - Studying causal mechanisms is "essential but challenging," and "triangulation of evidence is necessary."
  - Slides: `imai.fas.harvard.edu/talk-files/NBER25.pdf`
- **In practice:** always report a **sensitivity analysis** (not a bare "% mediated"); use **CDEs / sequential** g when you only need *whether* a direct path exists; triangulate (mediation, CDE, effect modification, placebo tests); and **manipulate the mediator** directly when you can.

Stay the Course

Star Wars IV. Obi-Wan gives Luke his father's lightsaber, the one Anakin always wanted him to have.

Source: https://snarkwars.com/2017/03/10/star-wars-episode-iv-part-ii-the-same-as-your-father/

Star Wars V. Yoda trains Luke on Dagobah, warning him away from the quick and easy path.



Source: imgflip.com/i/8alb2b

You've come a long way, but don't stop here. Keep reading papers, keep studying, keep questioning your own results. The quick and easy path is tempting, but you saw where it led Anakin.

Don’t forget to fill out your evaluations!

Darth Vader in Star Wars (1977) making a grand speech “May the force be with you.”

Source: https://www.theguardian.com/film/filmblog/2017/may/04/may-the-fourth-star-wars-best-lines

Onto the presentations & discussions!

Contact Information:
jaewon.yoo@iss.nthu.edu.tw
https://j1yoo.github.io/

(In)direct Effects with Non-binary Mediators

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

$$\overline{\delta}(d) = \sum_{\mathbf{x}} \Bigg(\sum_{m=0}^{J-1} \mathbb{E}[Y_i \mid M_i = m, D_i = d, \mathbf{X}_i = \mathbf{x}] \times \{\mathbb{P}[M_i = m \mid D_i = 1, \mathbf{X}_i = \mathbf{x}] - \mathbb{P}[M_i = m \mid D_i = 0, \mathbf{X}_i = \mathbf{x}]\} \Bigg) \times \mathbb{P}(\mathbf{X}_i = \mathbf{x})$$

- The ANDE is the following:

$$\overline{\zeta}(d) = \sum_{\mathbf{x}} \Bigg(\sum_{m=0}^{J-1} \{\mathbb{E}[Y_i \mid M_i = m, D_i = 1, \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i \mid M_i = m, D_i = 0, \mathbf{X}_i = \mathbf{x}]\} \times \mathbb{P}[M_i = m \mid D_i = d, \mathbf{X}_i = \mathbf{x}] \Bigg) \mathbb{P}(\mathbf{X}_i = \mathbf{x})$$

  - Effect of D_i for a fixed m averaged over the distribution of M_i when $D_i = d$.

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:

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

- 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.

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 $\mathbf{Z}_i$ which might be very high dimensional:

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

- 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.