Week 9: Mediation Analysis

PLSC 30600 - Causal Inference

Today

- How can we ask causal questions about **mechanisms**?
  - Does the effect of treatment go away if we intervene on some mediating variable?
  - What share of a treatment effect "flows" through a particular intermediate variable?
- Requires us to define *new* estimands
  - Consider "joint" interventions on the treatment and the mediator.
- Many challenges!
  - Assumptions for identifying mediation effects are not guaranteed even in a randomized experiment!

Mediation Estimands

Mechanisms

- Theory gives us beliefs not only about the existence of causal effects but the *reasons* for why these effects occur.
- Often, competing theories differ not in their predictions about the existence of an effect but rather the mechanisms through which they operate.
  - Ex. Brader, Valentino and Suhay (2008) - ethnicity-based framing effects of media on immigration on immigration attitudes
  - One theory argues that out-group vs. in-group cues about immigrants raise emotional anxiety which raises opposition to immigration
  - Another argues that these cues influence beliefs about economic costs about immigration.
- We want to articulate these competing theories in terms of statistical **estimands**
  - But what's the thought experiment?
- Mediation quantities consider the effect of a "joint" intervention on the treatment and the mediating variable.
  - e.g. Suppose we assign a unit to treatment, but fix its level of anxiety to a particular value
  - Suppose we assign a unit to treatment but fix its level of anxiety to **the level that would have been observed under control**

Mediation Effects

- Draw the DAG!
  - D_i is the treatment variable (binary)
  - M_i is the mediator
  - Define potential outcomes in terms of treatment and mediator $Y_i(d, m)$
- Consistency assumptions
  - $M_i = M_i(d)$ if $D_i = d$
  - $Y_i = Y_i(d, M_i(d))$ if $D_i = d$

Mediation effects

- Can think about different types of counterfactuals involving different joint interventions on treatment and mediator.
  - $Y_i(1, M_i(1))$ - The observed potential outcome under treatment.
  - $Y_i(1, 1)$ - The observed potential outcome under treatment where the mediator is fixed to 1
  - $Y_i(1, M_i(0))$ - The observed potential outcome under treatment where the mediator is set to the value it would take if i were assigned control.
- $Y_i(1, M_i(0))$ is a **cross-world** counterfactual - cannot be observed *at all*!

Controlled Direct Effects

- The Controlled Direct Effect considers a difference between intervening on treatment vs. control **also** fixing the mediator to a particular quantity

$$\text{CDE}(m) = Y_i(1, m) - Y_i(0, m)$$

- How to interpret
  - Effect of the treatment in the presence of some other intervention that fixed the mediator to m for all units
  - Imagine an experiment that manipulated both D_i and M_i
  - Can be used to show that there exists a path **unmediated** by M_i

Natural Indirect Effect

- The Natural Indirect Effect captures the change in the outcome if treatment is set fixed to d, but the mediator is set to the level it would take under treatment vs. the level it would take under control.

$$\text{NIE}(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$

- How to interpret
  - What is the effect of the treatment-induced change in the mediator, holding fixed the treatment itself.
- Vanderweele (2014) provides an interpretation in terms of actually observable potential outcomes

$$\text{NIE}(d) = \Big(Y_i(d, 1) - Y_i(d, 0)\Big) \times \Big(M_i(1) - M_i(0)\Big)$$

Natural Direct Effect

- The Natural Direct Effect differs slightly from the CDE since it imagines fixing d to $M_i(d)$ as opposed to m

$$\text{NDE}(d) = Y_i(1, M_i(d)) - Y_i(0, M_i(d))$$

- How to interpret:
  - What is the effect of treatment if each unit's mediator were set to the value it would take under d
  - Can conceptualize it as a "hypothetical" treatment that does not affect the mediator.
- With a binary mediator, Vanderweele (2014) provides a decomposition for in terms of a mixture of the two CDEs

$$\text{NDE}(d) = Y_i(1,0) - Y_i(0,0)) + \Big(Y_i(1,1) - Y_i(0,1) - Y_i(1,0) + Y_i(0,0)\Big) M_i(d)$$

- Under constant CDEs, the average controlled direct effect is the average natural direct effect

Effect Decomposition

- The goal of mediation analysis is typically to decompose the total effect of a treatment into components attributable and not attributable to the mediator.
- The classic 2-way decomposition from Robins and Greenland (1992):

$$Y_i(1) - Y_i(0) = \text{NDE}_i(d) + \text{NIE}_i(1-d)$$

- Vanderweele (2013) clarifies that $\text{NDE}(1)$ or $\text{NIE}(1)$ can be thought of as a "pure" direct/indirect effect plus an interaction. This leads to a 3-way decomposition (for a binary mediator)

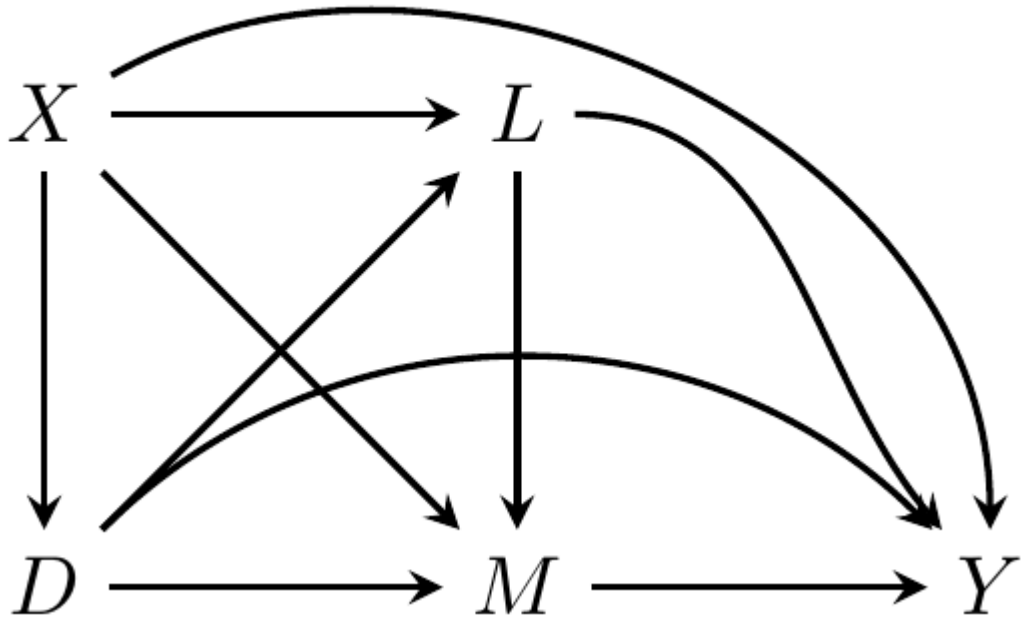
$$Y_i(1) - Y_i(0) = \text{NDE}_i(0) + \text{NIE}_i(0) + \Big(Y_i(1,1) - Y_i(0,1) - Y_i(1,0) + Y_i(0,0)\Big)\Big(M_i(1) - M_i(0)\Big)$$

- Writing the $NDE_i(0)$ in terms of the controlled direct effect fixing $M_i = 0$ leads to Vanderweele (2014)'s 4-way decomposition

$$\begin{aligned} Y_i(1) - Y_i(0) = \text{CDE}_i(0) + (Y_i(1,1) - Y_i(0,1) - Y_i(1,0) + Y_i(0,0))(M_i(d)) + \\ (Y_i(1,1) - Y_i(0,1) - Y_i(1,0) + Y_i(0,0))(M_i(1) - M_i(0)) + \text{NIE}_i(0) \end{aligned}$$

Identification

Identification of the ACDE



- Covariates
  - X_i: Pre-treatment confounders of treatment, mediator, and outcome.
  - L_i: Post-treatment confounders of mediator and outcome.

Identification of the ACDE

- Estimand: Average Controlled Direct Effect fixing $M_i = m$

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

- Identifying assumption: **Sequential ignorability**

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

- Treatment is as-good-as-randomly assigned given pre-treatment covariates
- Mediator is as-good-as-randomly assigned given pre-treatment covariates, post-treatment covariates and treatment.

Identification of the ACDE

- Can't just condition on L_i (affected by treatment) but can't *not* adjust for it (confounder of M).
- Solution: **Robins' g-formula**

$$E[Y_i(d,m)] = \sum_{x,l} E[Y_i|D_i = d, M_i = m, L_i = l, X_i = x] \times P(L_i = l|D_i = d, X_i = x) \times P(X_i = x)$$

- Challenges in direct estimation:
  - Need to model the distributions $P(L_i = l|D_i = d, X_i = x)$ and $P(X_i = x)$

Marginal Structural Models

- Robins (1999) developed a technique to estimate the average treatment effect of any *joint* intervention on multiple variables
  - Application to treatment history effects, etc...
  - Equally applicable here (since the CDE is just the effect of two non-randomized "treatments")
- First, define a "marginal structural model" for the mean potential outcomes under a particular treatment "history"

$$E[Y_i(d,m)] = \alpha_0 + \alpha_1 d + \alpha_2 m + \alpha_3 dm$$

- With long treatment histories need to make some assumptions (e.g. a blip + cumulative effect), but here straightforward to use a fully-saturated MSM.
- Estimation by IPTW.
  - Under sequential ignorability, we can get consistent estimates of the parameters via a regression with inverse-propensity of treatment weights.
  - For a unit with treatment $D_i = d$ and mediator $M_i = m$, we construct the IP weight:

$$\text{SW}_i = \frac{Pr(D_i = d)}{Pr(D_i = d|X_i = x)} \times \frac{Pr(M_i = m|D_i = d)}{Pr(M_i = m|D_i = d, X_i = x, L_i = l)}$$

- Now we just need two models - one for the treatment given X_i and another for the mediator given treatment, X_i and L_i.
- IPTW avoids the post-treatment bias problem of directly conditioning on L_i

Identification of the ANDE

- Identifying the average natural direct and indirect effects is a much greater challenge.
- Classic approaches from sociology relied on structural equations for the outcome and the mediator (Baron and Kenny)

$$
\begin{aligned}
Y_i &= \alpha_1 + \tau_1 D_i + + X_i' \beta_1 + \epsilon_{i1} \\
M_i &= \alpha_2 + \tau_2 D_i + X_i' \beta_2 + \epsilon_{i2} \\
Y_i &= \alpha_3 + \tau_3 D_i + \gamma M_i + X_i' \beta_3 + \epsilon_{i3}
\end{aligned}
$$

- Under what conditions can we get valid natural direct and indirect effects? Imai, Keele and Yamamoto (2010) show that we need a stronger version of **sequential ignorability** that rules out any intermediate confounders L_i

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

- For the SEM setting, we also need a no-interaction assumption: $\text{ANDE}(0) = \text{ANDE}(1)$ and the usual linearity assumptions.

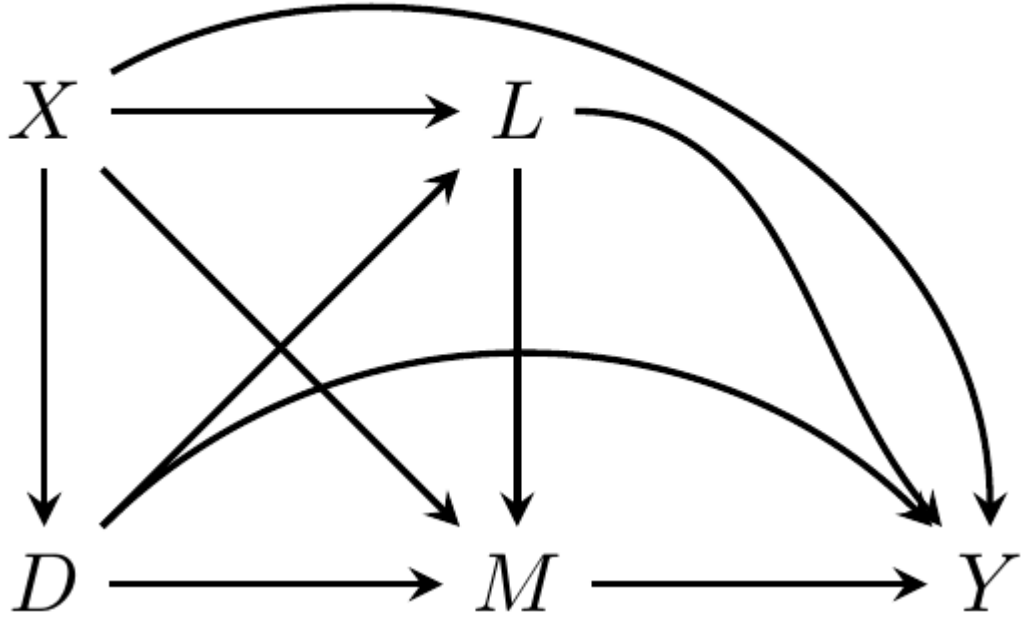
Identification of the ANDE

$$
\begin{aligned}
Y_i &= \alpha_1 + \tau_1 D_i + + X_i' \beta_1 + \epsilon_{i1} \\
M_i &= \alpha_2 + \tau_2 D_i + X_i' \beta_2 + \epsilon_{i2} \\
Y_i &= \alpha_3 + \tau_3 D_i + \gamma M_i + X_i' \beta_3 + \epsilon_{i3}
\end{aligned}
$$

- Under these assumptions, the classic "product of coefficients" method recovers the natural indirect effect
  - $\mathbf{TE} = \tau_1$
  - $\mathbf{ANIE}(0) = \mathbf{ANIE}(1) = \hat{\gamma}\hat{\tau_2}$
  - $\mathbf{ANDE}(0) = \mathbf{ANDE}(1) = \tau_1 - \hat{\gamma}\hat{\tau_2} = \hat{\tau_3}$
- Imai, Keele and Yamamoto (2010) show we can relax the no-interaction assumptions, but sequential ignorability without intermediate confounders is a key identifying assumption.

Estimating ACDEs with sequential G-estimation

Sequential g-Estimation



- **Sequential g-estimation** is an approach that makes the g-formula more tractable (and avoids having to model the density of the intermediate confounders) by making additional modeling assumptions on the impact of the mediator.
  - Part of a broader class of "structural nested mean models"
  - This specific approach was developed by Vansteelandt (2009) and Joffe and Greene (2009) and brought into political science by Acharya, Blackwell and Sen (2016)

Sequential g-Estimation

- Three step process

1. Run the regression of Y_i on everything

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i' \gamma_3 + L_i' \gamma_4 + \epsilon_i$$

1. Take $\hat{\gamma}_2$ as the effect of the mediator on the outcome and create a "blipped-down" outcome that removes the effect of the mediator on Y_i: $\tilde{Y}_i = Y_i - \hat{\gamma}_2 M_i$
2. Regress $\tilde{Y}_i$ on D_i and X_i only

$$\tilde{Y}_i = \beta_0 + \beta_1 D_i + X_i' \beta_3 + \epsilon_i$$

1. $\hat{\beta}_1$ is our estimate of the $\text{CDE}(0)$

Sequential g-Estimation

- **Additional assumptions**
  - In addition to sequential ignorability (conditional on intermediates), we need some additional modeling assumptions.
  - Obviously, an assumption that the linear model is correctly specified.
  - Also a "no-interaction" assumption between the effect of the mediator on the outcome and the intermediate covariates.

Illustration: Alesina, Giuliano and Nunn (2013)

- Research in "historical political economy" is interested in identifying the effects of the legacy of important historical factors on present-day phenomena.
  - A common challenge in HPE research is attributing a historical effect to a particular mechanism.
  - Critiques of HPE papers often have the form - "isn't this just driven by the effect of history on some intermediate X?"
  - Acharya, Blackwell and Sen (2016) argue that controlled direct effects are one such way of responding to these critiques. Non-zero CDEs are evidence that not all of an effect can be explained away by a particular mediating mechanism.
- Alesina, Giuliano and Nunn (2013) look at the impact of the adoption of plough-based agriculture on modern-day beliefs about gender equality.
  - Find evidence that plough agriculture leads to less equal gender norms and lower female labor-force participation.
  - Surprisingly find that there is no effect on the percent of women in political office.
  - One explanation: plough agriculture raises GDP per capita which contributes to more women in political office.
- What is the effect of plough agriculture *holding fixed* the intermediate variable of GDP per capita.
  - Most of our confounders of GDP per capita are *post* plough agriculture!

Illustration: Alesina, Giuliano and Nunn (2013)

- Y_i: Share of women in political positions in 2000
- D_i: Proportion of ethnic groups in a country that traditionally used the plough for agriculture
- M_i: log GDP per capita in 2000 (mean-centered)
- L_i: Post-treatment confounders
- X_i: Pre-treatment confounders (mostly geography)

Illustration: Alesina, Giuliano and Nunn (2013)

```
library(DirectEffects)
data(ploughs)

ate <- lm_robust(women_politics ~ plow + agricultural_suitability + tropical_climate + large_ar
tidy(ate) %>% filter(term == "plow")
```

```
##   term estimate std.error statistic p.value conf.low conf.high  df
## 1 plow     -2.1       2.1        -1   0.318    -6.25      2.04 145
##          outcome
## 1 women_politics
```

Illustration: Alesina, Giuliano and Nunn (2013)

- The current `DirectEffects` implementation of sequential-G specifies a formula as `y ~ d + x | l | m`
- Notably, if we think effects of the mediator might be non-linear, we can specify them as part of the model -- here we're assuming a quadratic relationship.

```
form_main <- 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

direct <- sequential_g(form_main, data = ploughs)
```

Illustration: Alesina, Giuliano and Nunn (2013)

```
summary(direct)
```

```
##
## t test of coefficients:
##
##                           Estimate Std. Err. t value Pr(>|t|)
## (Intercept)                 12.185     3.644    3.34   0.0011 **
## plow                        -4.839     2.345   -2.06   0.0413 *
## agricultural_suitability     4.574     3.105    1.47   0.1435
## tropical_climate            -2.189     2.105   -1.04   0.3006
## large_animals               -1.330     3.400   -0.39   0.6964
## political_hierarchies        0.496     1.091    0.45   0.6503
## economic_complexity         -0.105     0.430   -0.24   0.8070
## rugged                      -0.309     0.478   -0.65   0.5199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We estimate a CDE that's about 2x larger than the average treatment effect, consistent with our story about the income mediator masking a positive effect.
  - However, note that the difference between the average treatment effect and the controlled direct effect is *not* an indirect effect - as shown in Vanderweele (2014), it's a combination of mediation and interaction

Summary

- Empirically assessing mediation is a difficult problem

> So long as the limitations of this exploratory mode of investigation are clear, scientific investigation can proceed in an orderly manner. The problem is that so long as social scientists operate with a mistaken understanding of what can be expected from a mediation analysis, they will flit from one topic to another without an appropriate sense of the limits of what has been learned along the way. When critics make pious declarations about the importance of opening the black box, one must recognize that in social sciences black boxes are rarely if ever opened. Sometimes they are declared open by researchers who are too sanguine about the power of their lock-picking skills. Such declarations give the impression that the work is easy or already complete, which ironically slows the painstaking process by which real progress is made (Green, Ha and Bullock, 2010).

- Crucially: Assumptions for mediation effects cannot be guaranteed even in experiments where the treatment is randomized!
  - Designs *specifically* for mediation effects (e.g. Imai, Keele, Tingley and Yamamoto (2010)) rely on crossover exposures on the same individuals w/ a no-spillover assumption.
- Conceptually, interventions on a mediator variable may be much more poorly defined than interventions on treatment, making the estimands hard to interpret.