Mediation

William Lowe

Hertie School

4th November 2020

Mediation

Experiments give you effects, but *how* do those effects come about?

→ Mechanisms: operationalised as intervening / mediating variables

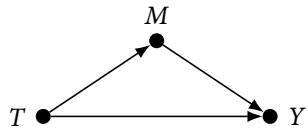
Mediation

Experiments give you effects, but *how* do those effects come about?

→ Mechanisms: operationalised as intervening / mediating variables

> *For example, the introduction of limes into the diet of seafarers in the 18th century dramatically reduced the incidence of scurvy, and eventually 20th century scientists figured out that the key mediating ingredient was vitamin C. Equipped with knowledge about why an experimental treatment works, scientists may devise other, possibly more efficient ways of achieving the same effect. Modern seafarers can prevent scurvy with limes or simply with vitamin C tablets.*
>
> *(Green et al., 2010)*

Mediation



The stylized structure of mediation. T affects Y in two ways

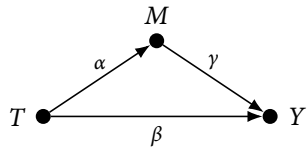
Indirect

- → the part of T's effect on Y that comes via its effect on M. (Note that M is observed)
- → represented by $T \longrightarrow M \longrightarrow Y$

Direct

- → *Every other* way T affects Y
- → Represented by $T \longrightarrow Y$

Mediation



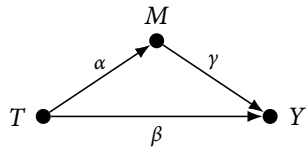
Old school (Baron & Kenny, 1986) mediation.

Everything is linear, conditionally Normal, decounfounded, and with three constant effects:

- → the direct effect: β
- → the indirect effect: $\alpha\gamma$
- → the total effect: $\alpha\gamma + \beta$

So far, so straightforward

Estimation



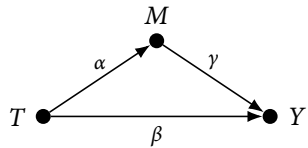
The *difference* approach: fit a model to get the total effect of T:

$$Y = a_0 + Ta_1 + \epsilon_a \qquad a_1 = \beta + \alpha\gamma$$

then a model controlling for M. Subtract treatment coefficients to get the indirect effect

$$Y = b_0 + Tb_1 + Mb_2 + \epsilon_b \qquad b_1 = \beta \qquad a_1 - b_1 = \alpha\gamma$$

Estimation



The *multiplicative* approach: fit a model to get the total effect of T:

$$Y = b_0 + Tb_1 + Mb_2 + \epsilon_b \qquad b_2 = \beta$$

then a model predicting M. Multiply coefficients to get the indirect effect

$$M = c_0 + Tc_1 + \epsilon_c \qquad c_1 b_2 = \alpha\gamma$$

Limitations

Does no work for non-linear systems, e.g. when M is a 'gate' that allows the direct effect to occur

Awkward model for, e.g. binary M

Not clear whether we can estimate everything with heterogeneity of effects

Effects

It's time to get counterfactual:

- → Stop talking about parameters
- → Start talking about effects

Express the fact that Y depends on T directly and indirectly via T's effect on M in (less than usually cumbersome) the potential outcome notation:

$$Y(T, M(T))$$

and define four treatment effects:

- → Total effect (TE)
- → Controlled direct effect (CDE)
- → Natural direct effect (NDE)
- → Natural indirect effect (NIE)

Effects

Let's start with the good old ATE

$$\text{TE: } E[Y(1, M(1)) - Y(0, M(0))]$$

intervene to set *everybody's* M to some value m

$$\text{CDE(m): } E[Y(1, m) - Y(0, m)]$$

everyone gets the M they would have when if untreated, but T varies

$$\text{NDE: } E[Y(1, M(0) - Y(0, M(0))]$$

no one is treated by everyone gets (or loses) the M that treatment would have given them

$$\text{NIE: } E[Y(0, M(1) - Y(0, M(0))] \qquad \text{NIE}_{\text{rev}} : E[Y(0, M(0) - Y(0, M(1))]$$

What use are these things?

The controlled direct effects are natural policy targets and can be identified by randomized experiments

The natural effects are fairly purely mechanical / counterfactual (mere randomization will not work)

> *Whereas the controlled direct effect is of interest when policy options exert control over values of variables (e.g., raising the level of a substance in patients' blood to a prespecified concentration), the natural direct effect is of interest when policy options enhance or weaken mechanisms or processes (e.g., freezing a substance at its current level of concentration [for each patient], but preventing it from responding to a given stimulus).*
>
> *(Pearl, 2014)*

Compare and contrast

In the simple linear systems we looked at before, many of these effects are the same

- → TE = $\beta + \alpha\gamma$
- → NDE = CDE(m) = β (i.e. when T does not interact with M)
- → NIE = $\alpha\gamma$

Note:

- → In linear systems TE = NDE + NIE
- → In non-linear ones, maybe not[1]

[1] TE = NDE - NIE_{rev}, but generally NIE = -NIE_{rev} only in linear systems

Estimating natural effects

In the absence of confounders:

$$\text{NDE} = \sum_m E[Y(1,m) - Y(0,m)]P(M = m \mid T = 0)$$

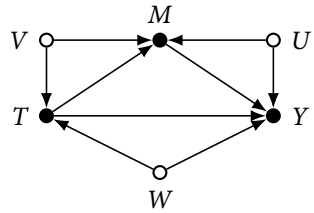
$$\text{NIE} = \sum_m E[Y(0,m)][P(M = m \mid T = 1) - P(M = m \mid T = 0)]$$

So the NDE is a average of CDE(m) weighted by the probability of each m value in the untreated population

and the NIE is an average of Y responses to the mediator, weighted by the mediator's responsiveness to treatment

→ This is a generalized analogue of the two coefficients in the multiplicative approach

Problems and assumptions

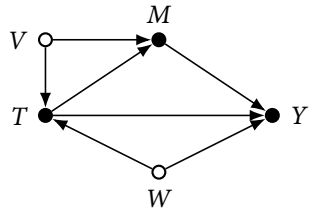


Four kinds of problems (and three kinds of assumptions)

- → Treatment outcome confounding: W
- → Treatment mediator confounding: M
- → Mediator outcome confounding: U
- → 'Intermediate confounding': $T \longrightarrow U$ (not show here for clarity)

Note: single variables may confound in multiple ways – we've just labeled them separately here

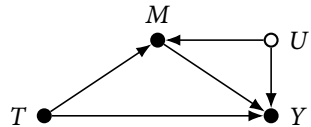
Familiar problems



Controlling this confounding is not different to a regular study

→ Randomize T
→ Measure and control for / weight / match on V and W

Problems



Confounding:

- → Randomising T does not deal with U confounding
- → Controlling for U will work
- → We might experimentally randomize M, but that will be a little weird… T causes M!

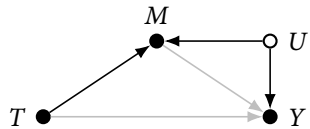
Confounded M means that the subtraction business doesn't work

- → M is a *collider* in the second regression

Note for later: U correlates the errors of a M on T regression (ϵ_c) and a Y on M and T regression (ϵ_b)

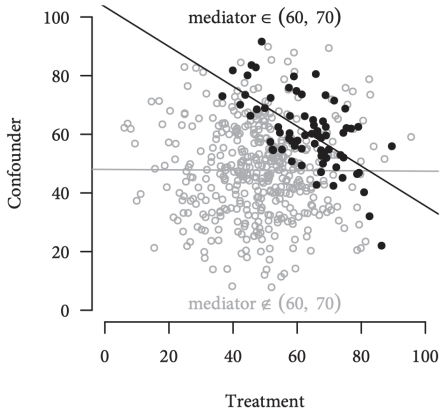
Problems

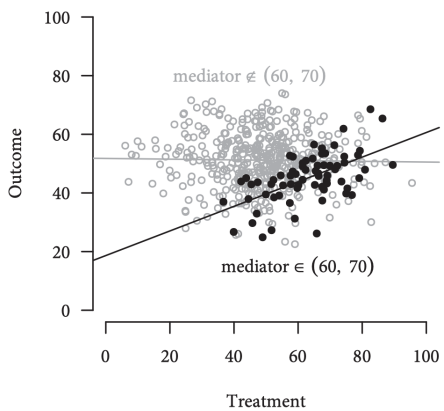
Consider an example where there is no direct effect and no indirect effect either



When we condition on M we'll get *pure collider bias*

Problems

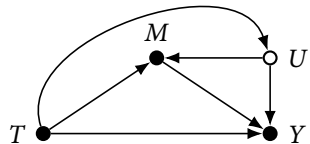




(Acharya et al., 2016, Fig. 2)

Problems

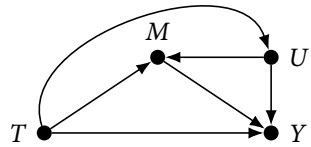
Some confounding relations are worse than others…Here treatment itself affects a mediator-outcome confounder



Acharya et al., 2016 example:

- → T ethnic fractionalisation
- → Y civil conflict
- → M political instability
- → U country GDP

Effects



With $T \longrightarrow U$

- → NDI and NIE aren't identified at all
- → CDE(m) is identified, but we can't get there with regressions alone

Why? Say we measure and then condition on U. This simultaneously removes

- → problematic confounding
- → some of the direct effect of T on Y!

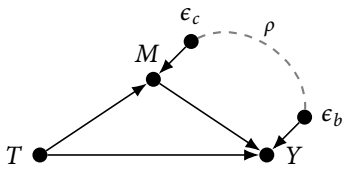
Sensitivity analysis

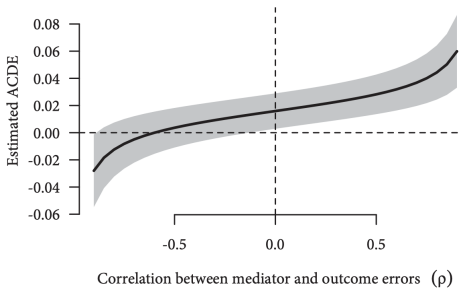
Lack of U confounding is a basic assumption for most mediation analysis

We can also ask:

→ How strong must the effect of U be for the CDE(m) to 'go away'

→ Assert ρ as the correlation between the M errors and Y errors





Mediation

How to be wrong about mechanisms

Mediation

How to be wrong about mechanisms

- → Limes aren't as good as lemons for preventing scurvy; not as much vitamin C
- → But the intuitive *M* was *acidity*, and limes are similar that way (also both species called citrus)
- → The British Navy introduced lemon rations after experimentation in the 18th century
- → Then rations improved and shipping times shortened
- → Then the Navy replaced with lemons with limes (but nobody noticed)
- → Scott raced Amundsen to the South Pole
- → and got scurvy again, in the 20th century…

Mediation

How to be wrong about mechanisms

- → Limes aren't as good as lemons for preventing scurvy; not as much vitamin C
- → But the intuitive *M* was *acidity*, and limes are similar that way (also both species called citrus)
- → The British Navy introduced lemon rations after experimentation in the 18th century
- → Then rations improved and shipping times shortened
- → Then the Navy replaced with lemons with limes (but nobody noticed)
- → Scott raced Amundsen to the South Pole
- → and got scurvy again, in the 20th century…

For the unfortunate but fascinating history of mediation analysis failure, see this [link] by Maciej Ceglowski.

References

Acharya, A., Blackwell, M. & Sen, M. (2016). 'Explaining causal findings without bias: Detecting and assessing direct effects'. *American Political Science Review*, *110*(03), 512–529.

Baron, R. M. & Kenny, D. A. (1986). 'The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations'. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Green, D. P., Ha, S. E. & Bullock, J. G. (2010). 'Enough already about 'black box' experiments: Studying mediation is more difficult than most scholars suppose'. *The Annals of the American Academy of Political and Social Science*, *628*(1), 200–208.

Pearl, J. (2014). 'Interpretation and identification of causal mediation'. *Psychological Methods*, *19*(4), 459–481.