Causal Mediation

Mark Andrews
Psychology Department, Nottingham Trent University

✉ mark.andrews@ntu.ac.uk

Establishing causal effects

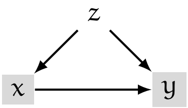
- We hypothesize that x causes y, and we aim to establish its causal effect.

$$x \longrightarrow y$$

- For example, y could be getting lung cancer, x could be smoking 20 cigarettes a day every day from an early age.
- For person i, $y_i \in \{0, 1\}$ is whether they get lung cancer and $x_i \in \{0, 1\}$ is whether they are a smoker (as just defined).

Confounds

- Of course, a third variable (or set of variables) z may also affect the value of x and y.



- For example, $z \in (-\infty, \infty)$ could be bad attitude towards healthy lifestyles generally. For example, if z is high, the person may have poor diet, lack of exercise, drink heavily, and smoke heavily.
- We will assume that we do not know z.

Potential outcomes framework

- In the *potential outcomes framework*, we consider what would have happened to any person i had they been *assigned* $x_i = 0$ or $x_i = 1$.
- If x_i was assigned the value of 0, then we say the potential outcomes of y_i is $y_i(0)$. Likewise, if x_i was assigned the value of 1, then we say the potential outcomes of y_i is $y_i(1)$.
- For person i, they always have two potential outcomes, $(y_i(0), y_i(1))$ and the causal effect of x_i is

$$\delta_i = y_i(1) - y_i(0).$$

- However, for any person i we observe

$$y_i = \begin{cases} y_i(0) & \text{if } x_i = 0, \\ y_i(1) & \text{if } x_i = 1. \end{cases}$$

- We only ever observe either $y_i(1)$ or $y_i(0)$ and never both, and so can never know δ_i. This is known as the *fundamental problem of causal inference*.

The power of randomization

- In observational studies, $(y_i(0), y_i(1))$ is not statistically independent of x_i.
- However, if x_i is assigned randomly, then $(y_i(0), y_i(1))$ *is* statistically independent of x_i.
- We can write this as follows:

$$(y_i(0), y_i(1)) \perp\!\!\!\perp x_i$$

which means

$$P(y_i(1)|x_i) = P(y_i(1)),$$
$$P(y_i(0)|x_i) = P(y_i(0)).$$

Average treatment effect (ATE)

- The *average treatment effect* (ATE) is the difference of average of the two potential outcomes.

$$
\begin{aligned}
\langle \delta \rangle &= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \delta_i, \\
&= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left(y_i(1) - y_i(0) \right), \\
&= \left(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} y_i(1) \right) - \left(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} y_i(0) \right), \\
&= \langle y(1) \rangle - \langle y(0) \rangle
\end{aligned}
$$

Average treatment effect under randomization

- When x_i is randomly assigned

$$\langle y(0) \rangle = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(y_i | x_i = 0) = \hat{y}_{[x_i=0]}$$

and

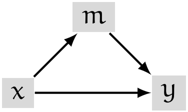
$$\langle y(1) \rangle = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(y_i | x_i = 1) = \hat{y}_{[x_i=1]}$$

- Therefore,

$$\langle y(1) \rangle - \langle y(0) \rangle = \hat{y}_{[x_i=1]} - \hat{y}_{[x_i=0]}$$

Mediation

- In a mediation model, we hypothesize that the effect of x on y is, at least partially, mediated by m.



- For example, to use a new example, attending a university (x) leads to high income (y) partially because it increases a person's social network (m).

Mediation and potential outcomes

- Assuming that x_i is binary, the potential values of the mediator are $(m_i(0), m_i(1))$.
- Thus, the potential outcomes are

$$y_i(0, m_i(0)), y_i(0, m_i(1)), y_i(1, m_i(0)), y_i(1, m_i(1)).$$

- For example, $y_i(1, m_i(0))$ is the value that y_i would take if $x_i = 1$ and m_i had the value it would take if $x_i = 0$.
- Using the concrete example, $y_i(1, m_i(0))$ is the income that person i would obtain had they been assigned to attend university, i.e. $x_i = 1$, but their social network is that which would have obtained had they not attended university.

Causal mediation, or indirect, effect

- The effect of the mediator when $x_i = 0$ is

$$\delta_i(0) = y_i(0, m_i(1)) - y_i(0, m_i(0))$$

which is the difference in the outcome that person i would have obtained had they not gone to university and m_i took the value of $m_i(1)$ rather than $m_i(0)$.

- The effect of the mediator when $x_i = 1$ is

$$\delta_i(1) = y_i(1, m_i(1)) - y_i(1, m_i(0))$$

which is the difference in the outcome that person i would have obtained had they gone to university and m_i took the value of $m_i(1)$ rather than $m_i(0)$.

- The *no-interaction* assumption is that

$$\delta_i = \delta_i(0) = \delta_i(1).$$

The direct effect

- The effect of the treatment when $m_i = m_i(0)$ is

$$\zeta_i(0) = y_i(1, m_i(0)) - y_i(0, m_i(0))$$

  which is the difference in the outcome that person i would have obtained had they gone to university compared to if they had not, assuming m_i took the value of $m_i(0)$.

- The effect of the treatment when $m_i = m_i(1)$ is

$$\zeta_i(1) = y_i(1, m_i(1)) - y_i(0, m_i(1))$$

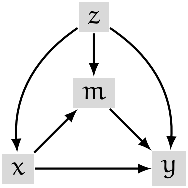
  which is the difference in the outcome that person i would have obtained had they gone to university compared to if they had not, assuming m_i took the value of $m_i(1)$.

- Again, the *no-interaction* assumption is that

$$\zeta_i = \zeta_i(0) = \zeta_i(1).$$

Confounders

- In general, there may be observed confounders z that affect x, m, and y.



- In our example, family background or intelligence etc might be examples of such confounders.

Sequential ignorability

- Sequential ignorability assumes:
  1. Potential outcomes and potential mediators are independent of treatment, conditional on covariates.
  2. Potential outcomes are independent of the mediators, conditional on the treatment.
- The first part holds if the treatment x is randomized.
- The second part does not necessarily even when x is randomized because the mediator is not randomized.

A general estimating algorithm

- Under sequential ignorability, we may estimate the average causal mediation effects (ACME) and average direct effect (ADE).
- Fit separate models outcome and mediator:

  1. Outcome model: $P(y_i|x_i, m_i, z_i)$.
  2. Mediator model: $P(m_i|x_i, z_i)$.

  where z_i are covariates.
- There is no restriction, e.g. no linearity, normality, etc, restrictions, on the nature of these regression models.

A general estimating algorithm

- Using the mediator model, predict $m_i(0)$ and $m_i(1)$ as follows:
  1. $m_i(0) = P(m_i | x_i = 0, z_i)$
  2. $m_i(1) = P(m_i | x_i = 1, z_i)$
- Using the outcome model, predict $y_i(1, m_i(0))$ and $y_i(1, m_i(1))$ as follows:
  1. $y_i(1, m_i(0)) = P(y_i | x_i = 1, m_i = m_i(0), z_i)$
  2. $y_i(1, m_i(1)) = P(y_i | x_i = 1, m_i = m_i(1), z_i)$
- Calculate $\hat{\delta}_i = y_i(1, m_i(1)) - y_i(1, m_i(0))$ and then

$$\hat{\delta} = \frac{1}{n} \sum_{i=1}^{n} \delta_i$$

is our estimator of δ, the average causal mediation effect (ACME).

A general estimating algorithm

- Using the outcome model, predict $y_i(1, m_i)$ and $y_i(0, m_i)$ as follows:
  1. $y_i(1, m_i) = P(y_i | x_i = 1, m_i, z_i)$
  2. $y_i(0, m_i) = P(y_i | x_i = 0, m_i, z_i)$
- Calculate $\hat{\zeta}_i = y_i(1, m_i) - y_i(0, m_i)$ and then

$$\hat{\zeta} = \frac{1}{n} \sum_{i=1}^{n} \zeta_i$$

  is our estimator of ζ, the average direct effect (ADE).