Causal Inference
CS 477-677

Mediation Analysis

Ilya Shpitser

JOHNS HOPKINS UNIVERSITY

Outline

1. Example Of Mediation
2. Treatment Decomposition
3. Identification
4. Estimation

Decomposing Causal Effects

- (Total) causal effects: $E[Y(1)] - E[Y(0)]$:
  - Randomize (Daniel 1-15, Lindt, Pierce, Neyman, Fisher):

$$\begin{tikzcd} & U \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd}$$

  - Observe confounders/stratify:

$$\begin{tikzcd} & U \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & Y \end{tikzcd}$$

- Parametric identification or bounds:
  - Instrumental variable + assumptions (P. Wright, 1928):

$$\begin{tikzcd} & & U \arrow[dl] \arrow[dr] & \\ Z \arrow[r] & A \arrow[rr] & & Y \end{tikzcd}$$

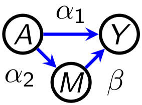
- Today: decomposing a causal effect into pathway components.

Motivating "Direct Effects": Discrimination

"The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same."

In Carson versus Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996).

Mediation Analysis (Linear Case)



- Total effect model:

$$Y = \alpha_0 A + \epsilon_0$$

- Mediation models:

$$Y = \alpha_1 A + \beta M + \epsilon_1$$
$$M = \alpha_2 A + \epsilon_2$$

- If above graph is true, $\alpha_0 =$ 'total effect", $\alpha_1 =$ "direct effect," $\alpha_2 \cdot \beta$ $=$ "indirect effect"
- Sewall Wright (1918) showed:

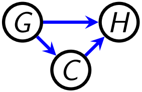
$$\text{total effect} = \text{direct effect} + \text{indirect effect}$$
$$\alpha_0 = \alpha_1 + \alpha_2 \cdot \beta$$

Direct And Indirect Effects With Arbitrary Models

- Cannot interpret coefficients causally in non-linear models.
- Want a general view, tying back to Hume/Lindt/experiments
- Will combine potential outcomes in a certain way to encode influence of A on Y along certain pathways.

Motivating "Direct Effects": Discrimination

- G (gender), C (characteristics), H (hiring).



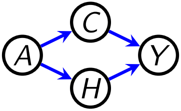
- Compare resumes of men:

$$H(G = male, C(G = male)) = H(G = male)$$

and *same* resumes with names switched to female ones:

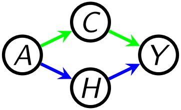
$$H(G = female, C(G = male)).$$

Motivating "Path-Specific Effects": Etiology by Pathway



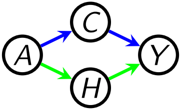
- A (smoking), C (cancer), H (heart disease), Y (outcome).
- A affects Y via smoke (C pathway), and via nicotine (H pathway).

Motivating "Path-Specific Effects": Etiology by Pathway



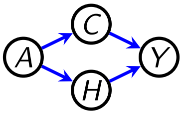
- A (smoking), C (cancer), H (heart disease), Y (outcome).
- A affects Y via smoke (C pathway), and via nicotine (H pathway).
- "Y if given nicotine-free cigarettes:" $Y(C(a), H(a'))$.

Motivating "Path-Specific Effects": Etiology by Pathway



- A (smoking), C (cancer), H (heart disease), Y (outcome).
- A affects Y via smoke (C pathway), and via nicotine (H pathway).
- "Y if given nicotine patches:" $Y(C(a'), H(a))$.

Motivating "Path-Specific Effects": Etiology by Pathway



- A (smoking), C (cancer), H (heart disease), Y (outcome).
- A affects Y via smoke (C pathway), and via nicotine (H pathway).
- "Y if given nicotine-free cigarettes:" $Y(C(a), H(a'))$.
- "Y if given nicotine patches:" $Y(C(a'), H(a))$.

Defining Direct Effects

- Total effect: $E[H(g=1)] - E[H(g=0)]$:

$$
\begin{tikzcd}
G \arrow[rr] \arrow[dr] & & H \\
& C \arrow[ur] &
\end{tikzcd}
$$

- Say G : male (0) vs female (1) name on resume, H is hiring decision (1 is yes, 0 is no).
- What is a sensible question for discrimination?
- Compare hiring based on male resumes with same resumes but with names switched to female:

$$E[H(g=0)] - E[H(g=1, C(g=0))]$$

- Or compare female resumes with same resumes but with names switched to male:

$$E[H(g=1)] - E[H(g=0, C(g=1))]$$

- These are called natural direct effects.

Defining Indirect Effects

- Non-zero direct effect corresponds to discrimination in this setting.
- Can also define indirect effect similarly.
- Compare hiring based on a woman's resume with a man's name vs hiring based on a man's resume and a man's name:

$$E[H(g=1, C(g=0))] - E[H(g=1, C(g=1))].$$

- Or (with genders switched):

$$E[H(g=0, C(g=1))] - E[H(g=0, C(g=0))].$$

Defining Indirect Effects

- Non-zero direct effect corresponds to discrimination in this setting.
- Can also define indirect effect similarly.
- Compare hiring based on a woman's resume with a man's name vs hiring based on a man's resume and a man's name:

$$E[H(g=1, C(g=0))] - E[H(g=1, C(g=1))].$$

- Or (with genders switched):

$$E[H(g=0, C(g=1))] - E[H(g=0, C(g=0))].$$

- We get the following decomposition (same with flipped genders):

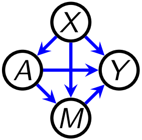
$$\underbrace{E[H(1)] - E[H(0)]}_{\text{ACE}} = \underbrace{(E[H(1)] - E[H(0, C(1))])}_{\text{Direct}} + \underbrace{(E[H(0, C(1))] - E[H(0)])}_{\text{Indirect}}$$

Nonparametric Effect Decomposition

- Defined direct and indirect effects and obtained a decomposition of the overall (total) causal effect.
- Did not mention statistical models (e.g. linear regressions) at all.
- Used potential outcomes directly.
- Can use any statistical model.
- But first, must make sure we are identified from observed data.

Simplest Interesting Mediation Setting

- $\vec{X}$ a vector of baseline factors/confounders (as before).
- A a treatment we are decomposing, M a mediator, Y an outcome.
- As before $\rightarrow$ means "directly causes."



- To get direct and indirect effects, need to identify the following three distributions: $p(Y(1)), p(Y(0)), p(Y(1, M(0)))$.
- Need assumptions.

Identifying Assumptions (Causal Model)

- Identifying Assumptions:
  - Conditional ignorability: $M(a) \perp\!\!\!\perp A \mid \vec{X}$ and $Y(a, m) \perp\!\!\!\perp \{A, M\} \mid \vec{X}$ for all a, m.
  - Means conditioning on $\vec{X}$ suffices to deal with confounding between A, M and Y, and between A and M.

Identifying Assumptions (Causal Model)

- Identifying Assumptions:
  - Conditional ignorability: $M(a) \perp\!\!\!\perp A \mid \vec{X}$ and $Y(a,m) \perp\!\!\!\perp \{A, M\} \mid \vec{X}$ for all a, m.
  - Means conditioning on $\vec{X}$ suffices to deal with confounding between A, M and Y, and between A and M.
  - $M(a') \perp\!\!\!\perp Y(m,a) \mid \vec{X}$ for all a, a'.
  - Means within levels of $\vec{X}$, causal mechanisms for M and Y have independent sources of noise, even if treatments "mismatch."

Identifying Assumptions (Causal Model)

- Identifying Assumptions:
  - Conditional ignorability: $M(a) \perp\!\!\!\perp A \mid \vec{X}$ and $Y(a, m) \perp\!\!\!\perp \{A, M\} \mid \vec{X}$ for all a, m.
  - Means conditioning on $\vec{X}$ suffices to deal with confounding between A, M and Y, and between A and M.
  - $M(a') \perp\!\!\!\perp Y(m, a) \mid \vec{X}$ for all a, a'.
  - Means within levels of $\vec{X}$, causal mechanisms for M and Y have independent sources of noise, even if treatments "mismatch."
  - Circuit analogy

On the board

Identifying Functionals

- Since we have conditional ignorability for Y and A,

$$p(Y(a)) = \sum_{\vec{X}} p(Y \mid A = a, \vec{X}) p(\vec{X}).$$

- Tricky case is $p(Y(a, M(a')))$:

Identifying Functionals

- Since we have conditional ignorability for Y and A,

$$p(Y(a)) = \sum_{\vec{X}} p(Y \mid A = a, \vec{X})p(\vec{X}).$$

- Tricky case is $p(Y(a, M(a')))$:

$$\begin{aligned}
p(Y(a, M(a'))) &= \sum_{m} p(Y(a, m), M(a') = m) \\
&= \sum_{m,\vec{X}} p(Y(a, m), M(a') = m \mid \vec{X})p(\vec{X}) \\
&= \sum_{m,\vec{X}} p(Y(a, m) \mid \vec{X})p(M(a') = m \mid \vec{X})p(\vec{X}) \\
&= \sum_{m,\vec{X}} p(Y(a, m) \mid a, m, \vec{X})p(M(a') = m \mid a', \vec{X})p(\vec{X}) \\
&= \sum p(Y \mid a, m, \vec{X})p(M = m \mid a', \vec{X})p(\vec{X})
\end{aligned}$$

Identifying Functionals

- Direct and indirect effects look like:

$$\sum_{\vec{X},m} \left\{ E[Y \mid A = 1, m, \vec{X}] - E[Y \mid A = 0, m, \vec{X}] \right\} p(m \mid A = 0, \vec{X}) p(\vec{X})$$

- How would we estimate this?
- With ACE either modeled Y or A.
- Have three models here: for Y, for M, and for A.
- Turns out any two are enough.

Parametric G-formula (Y and M Models)

- Model $E[Y \mid A, M, \vec{X}; \alpha]$, model $p(M \mid A, \vec{X}; \beta)$.
- Fit by MLE.
- Use empirical approximation for $p(\vec{X})$.
- If m is discrete, sum explicitly:

$$\frac{1}{n} \sum_{m,i} E[Y \mid a, m, \vec{X}_i; \hat{\alpha}] \cdot p(M = m \mid a', \vec{X}_i; \hat{\beta})$$

- If m is continuous, integrate by sampling:

$$\frac{1}{n} \int_m \sum_i E[Y \mid a, m, \vec{X}_i; \hat{\alpha}] \cdot p(M = m \mid a', \vec{X}_i; \hat{\beta})$$

- For certain parametric families for Y, M can do integral in closed form.

IPW (A and M Models)

- Model $p(A \mid \vec{X}; \gamma)$, model $p(M \mid A, \vec{X}; \beta)$.
- Fit by MLE.
- Use empirical approximation for $p(\vec{X})$.
- Remember, standard IPW gives us $E[Y(1)] = E[Y(1, M(1))]$:

$$\frac{1}{n} \sum_i \frac{\mathbb{I}(A = 1) Y_i}{p(A = 1 \mid \vec{X}_i; \hat{\gamma})}$$

- Want to somehow use M model to get mean of $E[Y(1, M(0))]$.

IPW (A and M Models)

- Model $p(A \mid \vec{X}; \gamma)$, model $p(M \mid A, \vec{X}; \beta)$.
- Fit by MLE.
- Use empirical approximation for $p(\vec{X})$.
- Remember, standard IPW gives us $E[Y(1)] = E[Y(1, M(1))]$:

$$\frac{1}{n}\sum_i \frac{\mathbb{I}(A=1)Y_i}{p(A=1 \mid \vec{X}_i; \hat{\gamma})}$$

- Want to somehow use M model to get mean of $E[Y(1, M(0))]$.
- Idea: just reweigh by the ratio:

$$\frac{1}{n}\sum_i \frac{\mathbb{I}(A=1)Y_i}{p(A=1 \mid \vec{X}_i; \hat{\gamma})} \cdot \frac{p(M \mid A=0, \vec{X}_i; \hat{\beta})}{p(M \mid A=1, \vec{X}_i; \hat{\beta})}.$$

Mixed Approach (Y and A Models)

- Model $p(A \mid \vec{X}; \gamma)$, model $E[Y \mid A, M, \vec{X}; \alpha]$.
- Fit by MLE.
- Use empirical approximation for $p(\vec{X})$.
- Remember, standard parametric g-formula gives us $E[Y(1)] = E[Y(1, M(1))]$:

$$\frac{1}{n} \sum_i E[Y \mid a, \vec{X}_i; \hat{\alpha}]$$

- Want to somehow use A mode to get mean of $E[Y(1, M(0))]$.

Mixed Approach (Y and A Models)

- Model $p(A \mid \vec{X}; \gamma)$, model $E[Y \mid A, M, \vec{X}; \alpha]$.
- Fit by MLE.
- Use empirical approximation for $p(\vec{X})$.
- Remember, standard parametric g-formula gives us $E[Y(1)] = E[Y(1, M(1))]$:

$$\frac{1}{n} \sum_i E[Y \mid a, \vec{X}_i; \hat{\alpha}]$$

- Want to somehow use A mode to get mean of $E[Y(1, M(0))]$.
- Idea: reweight observed 0 cases for M_i to get $M_i(0)$, but then use Y model with $A = 1$ and those M_i:

$$\frac{1}{n} \sum_i \frac{\mathbb{I}(A = 0)}{p(A = 0 \mid \vec{X}_i; \hat{\gamma})} \cdot E[Y \mid A = 1, m_i, \vec{X}_i; \hat{\alpha}]$$

Parametric G-Formula Pros and Cons

- Recall:

$$\frac{1}{n}\sum_{m,i} E[Y \mid a, m, \vec{X}_i; \hat{\alpha}] \cdot p(M = m \mid a', \vec{X}_i; \hat{\beta})$$

  - M, Y pro: most efficient use of data if models are right.
  - M, Y pro: fairly robust in practice.
  - M, Y con: easy to misspecify.
  - M, Y con: for continuous M, need to model density, have to integrate numerically.

IPW Pros and Cons

- Recall:

$$\frac{1}{n}\sum_i \frac{\mathbb{I}(A=1)Y_i}{p(A=1 \mid \vec{X}_i;\hat{\gamma})} \cdot \frac{p(M \mid A=0,\vec{X}_i;\hat{\beta})}{p(M \mid A=1,\vec{X}_i;\hat{\beta})}.$$

  - M, A pro: avoids modeling of complex Y.
  - M, A con: inefficient use of data.
  - M, A con: instability with small weights.
  - M, A con: for continuous M, need to model density, or model ratio directly (can you think of how?)

Mixed Approach Pros and Cons

- Recall:

$$\frac{1}{n}\sum_i \frac{\mathbb{I}(A=0)}{p(A=0 \mid \vec{X}_i; \hat{\gamma})} \cdot E[Y \mid A=1, M_i, \vec{X}_i; \hat{\alpha}]$$

  - Y, A pro: least amount of modeling: a mean and a binary probability.
  - Y, A con: somewhat inefficient use of data (but better than pure IPW).
  - M, A con: some instability with small weights (but better than pure IPW).

Testability

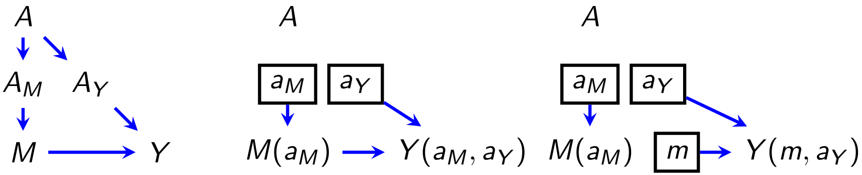
- Assumption $Y(a,m) \perp\!\!\!\perp M(a') \mid \vec{X}$ is kind of strange.
- Easy to test in circuits, not possible in people.
- Alternative is to explicitly split treatment (smoke/nicotine):

$$\begin{tikzcd} A \arrow[d] \arrow[dr] & \\ A' \arrow[d] & A'' \arrow[dr] & \\ M \arrow[rr] & & Y \end{tikzcd}$$

- Can check assumption via a hypothetical randomized trial that decomposes treatment.
- Alternative is sensitivity analysis (later).

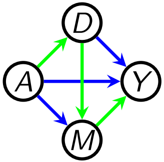
Representing Counterfactuals In Mediation Problems

- Recall, mediation as treatment decomposition
- A_M, A_Y are deterministic components of A associated with M, Y.
- Can intervene on them separately.
- Just split nodes as before (now into multiple pieces).
- a_M, a_Y potentially different.
- Read off independence by d-separation: $Y(m, a_Y) \perp\!\!\!\perp M(a_M)$.



Advanced Topic: Path-Specific Effects

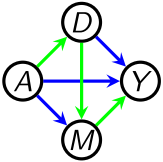
- Direct effect: along one arrow.
- Indirect effect: along all other arrows.
- Maybe we want effect along a specific path:



- Can do this by generalizing earlier idea.

Advanced Topic: Path-Specific Effects

- Direct effect: along one arrow.
- Indirect effect: along all other arrows.
- Maybe we want effect along a specific path:



- Can do this by generalizing earlier idea.

On the board

Next time: Representing Dependence, and Independence Using Graphs.