Mediation: Natural Direct and Indirect Effects

Ian Lundberg
Soc 212C2

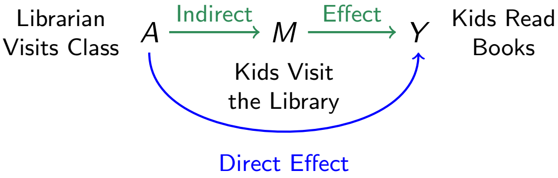
April 28, 2025

Learning goals for today

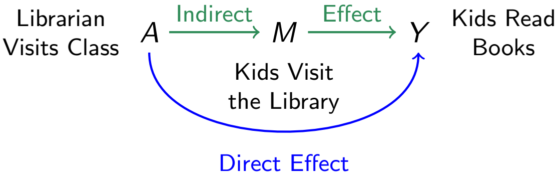
At the end of class, you will be able to:

1. Define natural direct and indirect effects
2. Decompose total effects into these components
3. Understand how intermediate confounding complicates natural direct effects
4. Estimate natural direct and indirect effects

Recall from before: Controlled direct effects



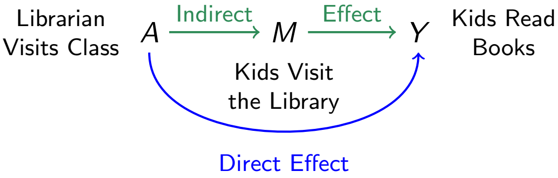
Recall from before: Controlled direct effects



Direct effect: Close the library

$$\mathrm{E}\left(Y^{10} - Y^{00}\right)$$

Recall from before: Controlled direct effects



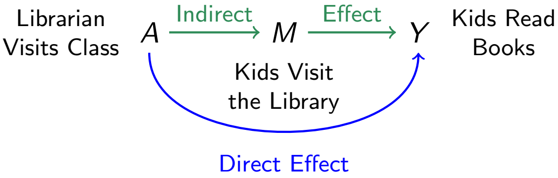
Direct effect: Close the library

$$\mathrm{E}\left(Y^{10} - Y^{00}\right)$$

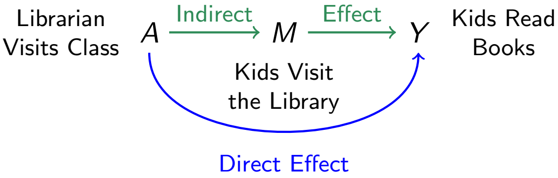
Direct effect: Make everyone visit the library

$$\mathrm{E}\left(Y^{11} - Y^{01}\right)$$

Today: Natural direct effects

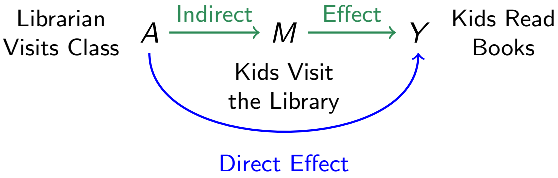


Today: Natural direct effects



Let M^a be the potential mediator value under treatment $A = a$

Today: Natural direct effects

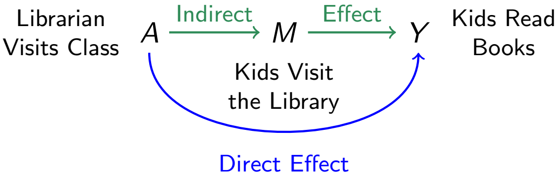


Let M^a be the potential mediator value under treatment $A = a$

Direct effect (0): Effect of the librarian visiting in a world where kids visit the library as if the librarian didn't visit the class

$$\mathrm{E}\left(Y^{1M^0} - Y^{0M^0}\right)$$

Today: Natural direct effects



Let M^a be the potential mediator value under treatment $A = a$

Direct effect (0): Effect of the librarian visiting in a world where kids visit the library as if the librarian didn't visit the class

$$\mathrm{E}\left(Y^{1M^0} - Y^{0M^0}\right)$$

Direct effect (1): Effect of the librarian visiting in a world where kids visit the library as if the librarian visited the class

$$\mathrm{E}\left(Y^{1M^1} - Y^{0M^1}\right)$$

Natural direct and indirect effects: Decomposition

The total effect can be decomposed into two components

Natural direct and indirect effects: Decomposition

The total effect can be decomposed into two components

$$\mathrm{E}\left(Y^1 - Y^0\right) \qquad \text{Total effect}$$

Natural direct and indirect effects: Decomposition

The total effect can be decomposed into two components

$$\mathrm{E}\Big(Y^1 - Y^0\Big) \qquad \text{Total effect}$$

$$= \mathrm{E}\Big(Y^{1M^1} - Y^{0M^0}\Big) \qquad \text{since } Y^a = Y^{aM^a}$$

Natural direct and indirect effects: Decomposition

The total effect can be decomposed into two components

$$\mathrm{E}\left(Y^1 - Y^0\right) \qquad \text{Total effect}$$

$$= \mathrm{E}\left(Y^{1M^1} - Y^{0M^0}\right) \qquad \text{since } Y^a = Y^{aM^a}$$

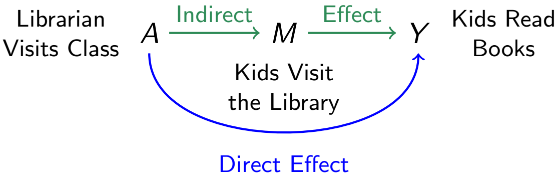
$$= \mathrm{E}\left(Y^{1M^1} \underbrace{- Y^{1M^0} + Y^{1M^0}}_{\text{Add } 0} - Y^{0M^0}\right)$$

Natural direct and indirect effects: Decomposition

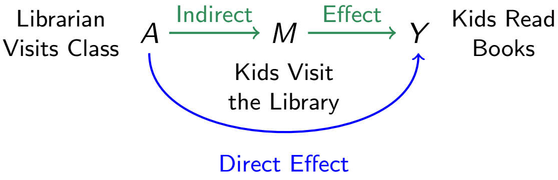
The total effect can be decomposed into two components

$$
\begin{aligned}
& E\left(Y^1 - Y^0\right) && \text{Total effect} \\
&= E\left(Y^{1M^1} - Y^{0M^0}\right) && \text{since } Y^a = Y^{aM^a} \\
&= E\left(Y^{1M^1} \underbrace{-Y^{1M^0} + Y^{1M^0}}_{\text{Add 0}} - Y^{0M^0}\right) \\
&= \underbrace{E\left(Y^{1M^1} - Y^{1M^0}\right)}_{\substack{\textbf{Indirect Effect} \\ \text{(effect through } M\text{)}}} + \underbrace{E\left(Y^{1M^0} - Y^{0M^0}\right)}_{\substack{\textbf{Direct Effect} \\ \text{(effect not through } M\text{)}}}
\end{aligned}
$$

Indirect effect: Interpretation



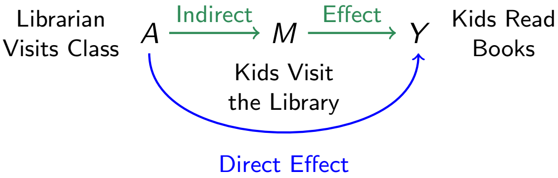
Indirect effect: Interpretation



$$\tau(1) = \mathsf{E}(Y^{1M^1} - Y^{1M^0})$$

$$\tau(0) = \mathsf{E}(Y^{0M^1} - Y^{0M^0})$$

Indirect effect: Interpretation

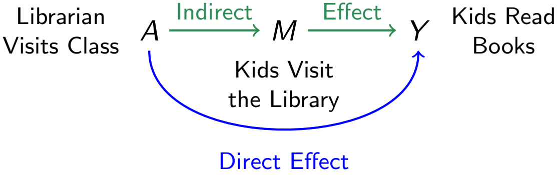


Indirect effect (1): Effect of visiting the library as much as you would if the librarian did vs did not visit,
in a world where the librarian visits

$$\tau(1) = \mathrm{E}(Y^{1M^1} - Y^{1M^0})$$

$$\tau(0) = \mathrm{E}(Y^{0M^1} - Y^{0M^0})$$

Indirect effect: Interpretation



Indirect effect (1): Effect of visiting the library as much as you would if the librarian did vs did not visit,
in a world where the librarian visits

$$\tau(1) = \mathrm{E}(Y^{1M^1} - Y^{1M^0})$$

Indirect effect (0): Effect of visiting the library as much as you would if the librarian did vs did not visit,
in a world where the librarian does not visit

$$\tau(0) = \mathrm{E}(Y^{0M^1} - Y^{0M^0})$$

Controlled and natural direct effects: Context of controlled experiments

Controlled direct effect

$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{11} - Y_i^{01}\right)$

Natural direct effect

$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{1M_i^1} - Y_i^{0M_i^1}\right)$

Controlled and natural direct effects: Context of controlled experiments

Controlled direct effect

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{11} - Y_i^{01}\right)$$

Y_i^{11} can be observed
in an experiment
— Assign $A = 1$ and $M = 1$

Natural direct effect

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{1M_i^1} - Y_i^{0M_i^1}\right)$$

Controlled and natural direct effects: Context of controlled experiments

Controlled direct effect

$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{11} - Y_i^{01}\right)$

Y_i^{11} can be observed in an experiment

— Assign $A = 1$ and $M = 1$

Y_i^{01} can be observed in an experiment

— Assign $A = 0$ and $M = 1$

Natural direct effect

$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{1M_i^1} - Y_i^{0M_i^1}\right)$

Controlled and natural direct effects: Context of controlled experiments

Controlled direct effect

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{11} - Y_i^{01}\right)$$

Y_i^{11} can be observed in an experiment
— Assign $A = 1$ and $M = 1$

Y_i^{01} can be observed in an experiment
— Assign $A = 0$ and $M = 1$

Natural direct effect

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{1M_i^1} - Y_i^{0M_i^1}\right)$$

$Y_i^{1M_i^1}$ can be observed in an experiment
— Assign $A = 1$

Controlled and natural direct effects: Context of controlled experiments

Controlled direct effect

$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{11} - Y_i^{01}\right)$

Y_i^{11} can be observed
in an experiment
— Assign $A = 1$ and $M = 1$

Y_i^{01} can be observed
in an experiment
— Assign $A = 0$ and $M = 1$

Natural direct effect

$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{1M_i^1} - Y_i^{0M_i^1}\right)$

$Y_i^{1M_i^1}$ can be observed
in an experiment
— Assign $A = 1$

$Y_i^{0M_i^1}$ **cannot** be observed
in an experiment
— Because M_i^1 is unknown!

The unobservable potential outcome: An experimental solution

[1] Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(1), 5-51.

The unobservable potential outcome: An experimental solution

Crossover design[1]

[1] Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(1), 5-51.

The unobservable potential outcome: An experimental solution

Crossover design[1]

- In period 1, assign $A = a$. See outcome $M = M^a$

[1] Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(1), 5-51.

The unobservable potential outcome: An experimental solution

Crossover design[1]

- In period 1, assign $A = a$. See outcome $M = M^a$
- In period 2, assign $A = a'$. Assign $M = M^a$. See $Y^{a' M^a}$

[1] Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(1), 5-51.

The unobservable potential outcome: An experimental solution

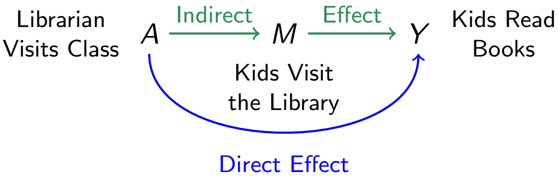
Crossover design[1]

- In period 1, assign $A = a$. See outcome $M = M^a$
- In period 2, assign $A = a'$. Assign $M = M^a$. See $Y^{a' M^a}$

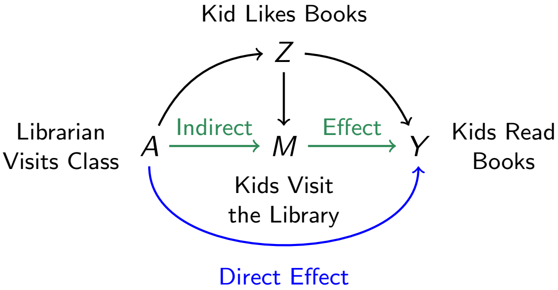
Works an assumption of no carry-over:
current treatment is all that affects current outcome

[1] Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(1), 5-51.

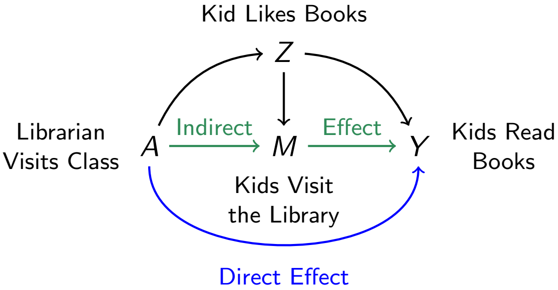
Intermediate confounding: A threat to natural effects



Intermediate confounding: A threat to natural effects



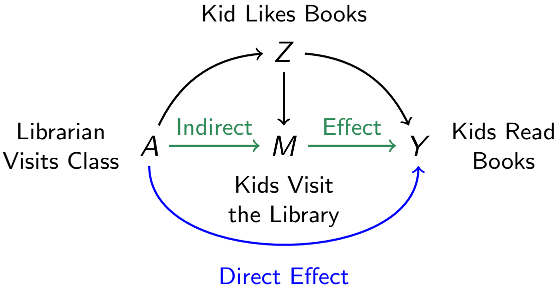
Intermediate confounding: A threat to natural effects



If Z is measured,

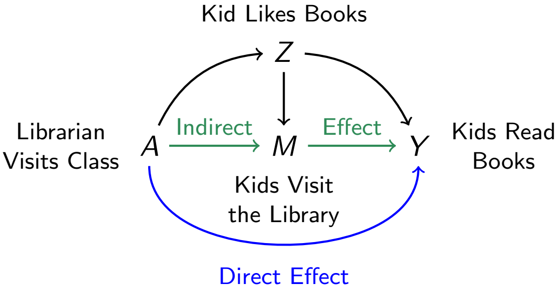
- The controlled direct effect $\text{E}(Y^{1a} - Y^{0a})$ is nonparametrically identified
- but the natural direct effect $\text{E}(Y^{1M^a} - Y^{0M^a})$ is not

Intermediate confounding: A threat to natural effects



Why?

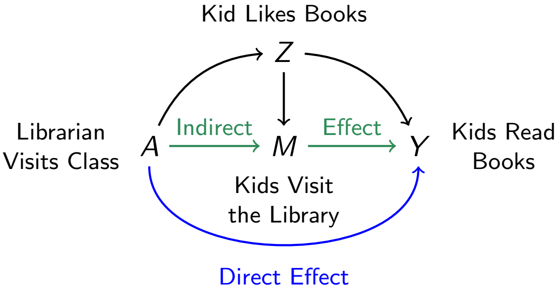
Intermediate confounding: A threat to natural effects



Why? For CDE:

$$E(Y^{10}) = E_{Z|A=1}\left(E\left(Y \mid A=1, M=0, Z\right)\right)$$

Intermediate confounding: A threat to natural effects



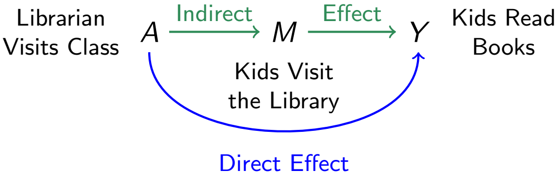
Why? For NDE:

$$\mathsf{E}(Y^{1M^0}) = \mathsf{E}_{Z|A=1}\left(\mathsf{E}\left(Y \mid A=1, M=M^0, Z\right)\right)$$

but we never have $M = M^0$ with $A = 1$, so that doesn't work

Intermediate confounding: A threat to natural effects

For NDE, this setting is nonparametrically identified



Intermediate confounding: A threat to natural effects

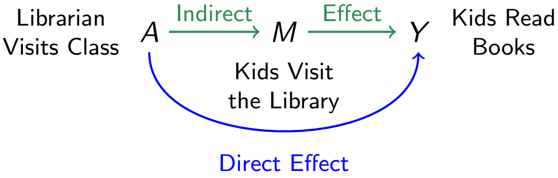
For NDE, this setting is nonparametrically identified

$$\begin{tikzcd}[column sep=large] \text{Librarian Visits Class} \quad A \arrow[r, "\text{Indirect}"] \arrow[rr, bend right=60, "\text{Direct Effect}"'] & \underset{\text{Kids Visit the Library}}{M} \arrow[r, "\text{Effect}"] & Y \quad \text{Kids Read Books} \end{tikzcd}$$

$\mathrm{E}(Y^{1M^0})$

Intermediate confounding: A threat to natural effects

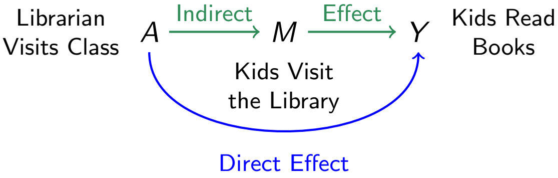
For NDE, this setting is nonparametrically identified



$$\mathrm{E}(Y^{1M^0}) = \mathrm{E}\left(Y \mid A = 1, M = M^0\right)$$

Intermediate confounding: A threat to natural effects

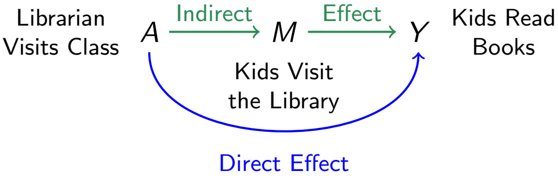
For NDE, this setting is nonparametrically identified



$$
\begin{aligned}
\mathrm{E}(Y^{1M^0}) &= \mathrm{E}\left(Y \mid A = 1, M = M^0\right) \\
&= \mathrm{P}(M^0 = 1)\mathrm{E}\left(Y \mid A = 1, M = 1\right)
\end{aligned}
$$

Intermediate confounding: A threat to natural effects

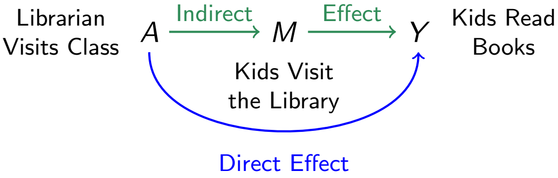
For NDE, this setting is nonparametrically identified



$$
\begin{aligned}
\mathrm{E}(Y^{1M^0}) &= \mathrm{E}\left(Y \mid A = 1, M = M^0\right) \\
&= \mathrm{P}(M^0 = 1)\mathrm{E}\left(Y \mid A = 1, M = 1\right) \\
&\qquad + \mathrm{P}(M^0 = 0)\mathrm{E}\left(Y \mid A = 1, M = 0\right) \\
&= \mathrm{P}(M = 1 \mid A = 0)\mathrm{E}\left(Y \mid A = 1, M = 1\right)
\end{aligned}
$$

Intermediate confounding: A threat to natural effects

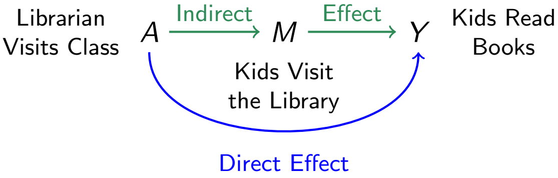
For NDE, this setting is nonparametrically identified



$$
\begin{aligned}
\mathrm{E}(Y^{1M^0}) &= \mathrm{E}\left(Y \mid A = 1, M = M^0\right) \\
&= \mathrm{P}(M^0 = 1)\mathrm{E}\left(Y \mid A = 1, M = 1\right) \\
&\qquad + \mathrm{P}(M^0 = 0)\mathrm{E}\left(Y \mid A = 1, M = 0\right) \\
&= \mathrm{P}(M = 1 \mid A = 0)\mathrm{E}\left(Y \mid A = 1, M = 1\right) \\
&\qquad + \mathrm{P}(M = 0 \mid A = 0)\mathrm{E}\left(Y \mid A = 1, M = 0\right)
\end{aligned}
$$

Intermediate confounding: A threat to natural effects

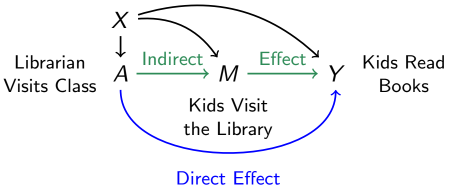
For NDE, this setting is nonparametrically identified



$$
\begin{aligned}
\mathrm{E}(Y^{1M^0}) &= \mathrm{E}\left(Y \mid A = 1, M = M^0\right) \\
&= \mathrm{P}(M^0 = 1)\mathrm{E}\left(Y \mid A = 1, M = 1\right) \\
&\qquad + \mathrm{P}(M^0 = 0)\mathrm{E}\left(Y \mid A = 1, M = 0\right) \\
&= \mathrm{P}(M = 1 \mid A = 0)\mathrm{E}\left(Y \mid A = 1, M = 1\right) \\
&\qquad + \mathrm{P}(M = 0 \mid A = 0)\mathrm{E}\left(Y \mid A = 1, M = 0\right)
\end{aligned}
$$

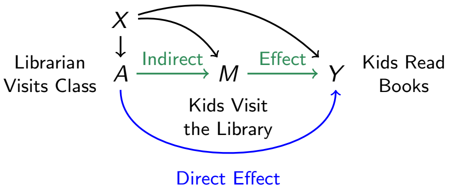
Identified because no potential outcomes remain!

Estimation for NDE when sequential ignorability holds[2]



[2]Method implemented in the R package `mediation` and described on p. 773: Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. American Political Science Review, 105(4), 765-789.

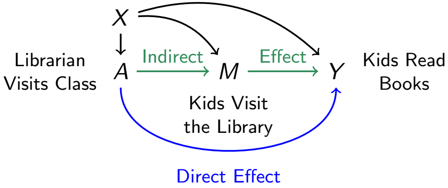
Estimation for NDE when sequential ignorability holds[2]



1. Model $\text{E}(M \mid X, A)$

[2]Method implemented in the R package `mediation` and described on p. 773: Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. American Political Science Review, 105(4), 765-789.

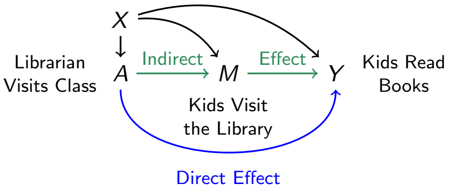
Estimation for NDE when sequential ignorability holds[2]



1. Model $\mathrm{E}(M \mid X, A)$
   - Predict M^a for all a

[2]Method implemented in the R package `mediation` and described on p. 773: Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. American Political Science Review, 105(4), 765-789.

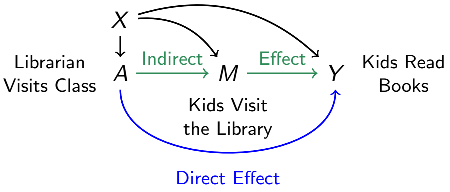
Estimation for NDE when sequential ignorability holds[2]



1. Model $\text{E}(M \mid X, A)$
   - Predict M^a for all a
2. Model $\text{E}(Y \mid X, A, M)$

[2]Method implemented in the R package `mediation` and described on p. 773: Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. American Political Science Review, 105(4), 765-789.

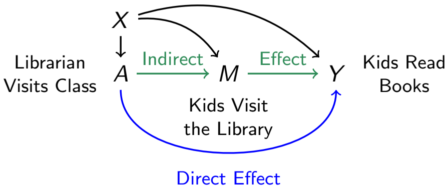
Estimation for NDE when sequential ignorability holds[2]



1. Model $\text{E}(M \mid X, A)$
   - Predict M^a for all a
2. Model $\text{E}(Y \mid X, A, M)$
   - Predict Y^{a', M^a} for any pair a', a

[2]Method implemented in the R package `mediation` and described on p. 773: Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. American Political Science Review, 105(4), 765-789.

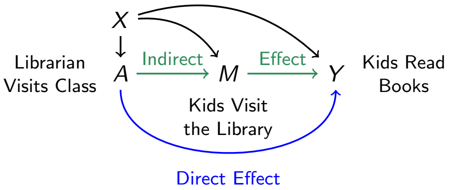
Estimation for NDE when sequential ignorability holds[2]



1. Model $\mathrm{E}(M \mid X, A)$
   - Predict M^a for all a
2. Model $\mathrm{E}(Y \mid X, A, M)$
   - Predict Y^{a', M^a} for any pair a', a
3. Average over the sample.

[2]Method implemented in the R package `mediation` and described on p. 773: Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. American Political Science Review, 105(4), 765-789.

Estimation for NDE when sequential ignorability holds[2]



1. Model $\text{E}(M \mid X, A)$
   - Predict M^a for all a
2. Model $\text{E}(Y \mid X, A, M)$
   - Predict Y^{a', M^a} for any pair a', a
3. Average over the sample.

Bootstrap for confidence intervals

[2] Method implemented in the R package `mediation` and described on p. 773: Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. American Political Science Review, 105(4), 765-789.

Summary: Mediation decomposes causal effects

Controlled direct effects — Example: $E(Y^{10} - Y^{00})$

- Idea: Intervene to hold the mediator at a fixed value
- ✓ Identified in a sequentially randomized experiment
- ✓ Identifiable with observed intermediate confounding
- **X** Direct and indirect effects are not additively decomposable

Natural direct and indirect effects — Example: $E(Y^{1M^0} - Y^{0M^0})$

- Idea: Intervene to hold the mediator at the value in the absence of treatment
- **X** Not identified[3] in an experiment—Y^{1M^0} is unobservable
  - Crossover experiments can help with an extra assumption
- **X** Not identifiable with any intermediate confounding
- ✓ Direct and indirect effects are additively decomposable

[3] Each use of "identified" refers to nonparametric identification. Parametric identification is sometimes possible when nonparametric identification is not.

Learning goals for today

At the end of class, you will be able to:

1. Define natural direct and indirect effects
2. Decompose total effects into these components
3. Understand how intermediate confounding complicates natural direct effects
4. Estimate natural direct and indirect effects

SUPPLEMENTAL

Natural direct and indirect effects in SWIGs

Direct effect under mediator $M = M^1$

$$\begin{tikzcd} A \mid 1 \arrow[r] \arrow[rr, bend left] & M \mid M^1 \arrow[r] & Y \end{tikzcd} \quad - \quad \begin{tikzcd} A \mid 0 \arrow[r] \arrow[rr, bend left] & M \mid M^1 \arrow[r] & Y \end{tikzcd}$$

Direct effect under mediator $M = M^0$

$$\begin{tikzcd} A \mid 1 \arrow[r] \arrow[rr, bend left] & M \mid M^0 \arrow[r] & Y \end{tikzcd} \quad - \quad \begin{tikzcd} A \mid 0 \arrow[r] \arrow[rr, bend left] & M \mid M^0 \arrow[r] & Y \end{tikzcd}$$

Indirect effect under treatment $A = 1$

$$\begin{tikzcd} A \mid 1 \arrow[r] \arrow[rr, bend left] & M \mid M^1 \arrow[r] & Y \end{tikzcd} \quad - \quad \begin{tikzcd} A \mid 1 \arrow[r] \arrow[rr, bend left] & M \mid M^0 \arrow[r] & Y \end{tikzcd}$$

Indirect effect under treatment $A = 0$

$$\begin{tikzcd} A \mid 0 \arrow[r] \arrow[rr, bend left] & M \mid M^1 \arrow[r] & Y \end{tikzcd} \quad - \quad \begin{tikzcd} A \mid 0 \arrow[r] \arrow[rr, bend left] & M \mid M^0 \arrow[r] & Y \end{tikzcd}$$