Causal Mechanisms

Kosuke Imai

Harvard University

Spring 2021

Causal Mechanisms

- Causal inference is a central goal of scientific research
- Scientists care about causal mechanisms, not just about causal effects $\rightsquigarrow$ external validity
- Policy makers want to devise better policies

- Randomized experiments often only determine whether the treatment causes changes in the outcome
- Not how and why the treatment affects the outcome
- Common criticism of experiments and statistics:

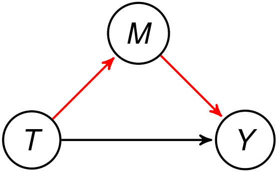
  black box view of causality

- Qualitative research $\rightsquigarrow$ process tracing

- Question: How can we learn about causal mechanisms from experimental and observational studies?

Direct and Indirect Effects

- DAG representation
  - $T \in \{0,1\}$: treatment
  - $M \in \mathcal{M}$: mediator
  - Y: outcome with potential outcome $Y(t, m)$



- Goal: decompose total effect into direct and indirect effects
- Alternative: decompose the treatment into different components
- How large is the indirect effect relative to the total effect?

Controlled Direct Effects (CDE)

- Definition

$$\begin{aligned} \text{Individual:} \quad \xi_i(m) &= Y_i(1, m) - Y_i(0, m) \\ \text{Average:} \quad \bar{\xi}(m) &= \mathbb{E}\{Y_i(1, m) - Y_i(0, m)\} \end{aligned}$$

  for some $m \in \mathcal{M}$

- Interpretation
  - direct effect of treatment while holding the mediator constant at m
  - causal effect of intervention on T and M

- CDE does not directly quantify causal mechanism
- If M fully explains causal mechanism, CDEs will be zero for all m
- Interaction effects $\xi_i(m) \neq \xi_i(m')$: CDE varies as a function of M

Natural Indirect Effects (NIE)

- Definition (Causal mediation effects)

$$\begin{aligned} \text{Individual:} \quad \delta_i(t) &= Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \\ \text{Average (ACME):} \quad \bar{\delta}(t) &= \mathbb{E}\{Y_i(t, M_i(1)) - Y_i(t, M_i(0))\} \end{aligned}$$

- Interpretation
  - effect of the change in M on Y that would be induced by T
  - change M from $M_i(0)$ to $M_i(1)$ while holding T at $t = 0$ or $t = 1$
  - zero treatment effect on M $\rightsquigarrow$ zero causal mediation effect

- Represents the causal mechanism through M_i
- Allows for the decomposition of treatment effect into direct and indirect effects

Treatment Effect Decomposition

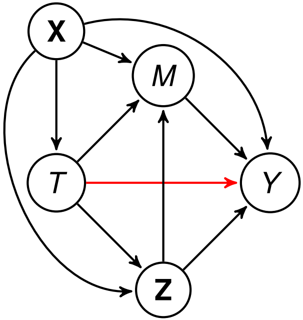
- Natural direct effect (NDE):

$$\begin{aligned} \text{Indiviual:} \quad \zeta_i(t) &= Y_i(1, M_i(t)) - Y_i(0, M_i(t)) \\ \text{Average:} \quad \bar{\zeta}(t) &= \mathbb{E}\{Y_i(1, M_i(t)) - Y_i(0, M_i(t))\} \end{aligned}$$

  - change T from 0 to 1 while holding M constant at $M_i(t)$
  - causal effect of T on Y, holding M constant at its potential value that would be realized when $T_i = t$
- Represents all mechanisms other than through M
- Effect decomposition:

$$\begin{aligned} \underbrace{Y_i(1, M_i(1)) - Y_i(0, M_i(0))}_{\text{total effect}} &= \underbrace{\delta_i(t)}_{\text{NIE}} + \underbrace{\zeta_i(1-t)}_{\text{NDE}} \\ &= \frac{1}{2}\sum_{t=0}^{1}\{(\delta_i(t) + \zeta_i(t))\} \end{aligned}$$

Identification of Controlled Direct Effects



- $\mathbf{X}$: pre-treatment confounders
- $\mathbf{Z}$: post-treatment confounders
- Assumptions:

$$\{Y_i(t,m), M_i(t')\} \perp\!\!\!\perp T_i \mid \mathbf{X}_i = \mathbf{x}$$
$$Y_i(t,m) \perp\!\!\!\perp M_i \mid \mathbf{X}_i = \mathbf{x}, T_i = t, \mathbf{Z}_i = \mathbf{z}$$

for all $t, \mathbf{x}, \mathbf{z}$

- Post-treatment bias: cannot simply "control for" $\mathbf{Z}$

$$\bar{\xi}(m) \neq \sum_{\mathbf{x},\mathbf{z}} \{\mathbb{E}(Y \mid T=1, M=m, \mathbf{X}, \mathbf{Z}) - \mathbb{E}(Y \mid T=0, M=m, \mathbf{X}, \mathbf{Z})\} P(\mathbf{X}, \mathbf{Z})$$

- Identification: must model $\mathbf{Z}$ given T and $\mathbf{X}$

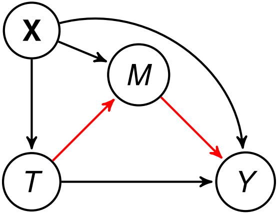
$$\bar{\xi}(m) = \sum_{\mathbf{x},\mathbf{z}} \{\mathbb{E}(Y \mid T=1, M=m, \mathbf{X}, \mathbf{Z}) P(\mathbf{Z} \mid T=1, \mathbf{X})$$
$$- \mathbb{E}(Y \mid T=0, M=m, \mathbf{X}, \mathbf{Z}) P(\mathbf{Z} \mid T=0, \mathbf{X})\} \, P(\mathbf{X})$$

Identification of Natural Direct and Indirect Effects

- No post-treatment confounders
- Assumptions:

$$\{Y_i(t,m), M_i(t')\} \perp\!\!\!\perp T_i \mid \mathbf{X}_i = \mathbf{x}$$
$$Y_i(t',m) \perp\!\!\!\perp M_i(t) \mid \mathbf{X}_i = \mathbf{x}, T_i = t$$

for all $t, t', \mathbf{x}$



- Cross-world counterfactuals
- Randomization of T, M does not satisfy the assumption
- Identification

$$\bar{\delta}(t) = \sum_{M,\mathbf{X}} \mathbb{E}(Y \mid M, T=t, \mathbf{X}) \{P(M \mid T=1, \mathbf{X}) - P(M \mid T=0, \mathbf{X})\} P(\mathbf{X})$$

$$\bar{\zeta}(t) = \sum_{M,\mathbf{X}} \{\mathbb{E}(Y \mid M, T=1, \mathbf{X}) - \mathbb{E}(Y \mid M, T=0, \mathbf{X})\} \times P(M \mid T=t, \mathbf{X}) P(\mathbf{X})$$

Estimation of Controlled Direct Effects

1. Directly use the identification formula

$$\bar{\xi}(m) = \sum_{\mathbf{X},\mathbf{Z}} \{\mathbb{E}(Y \mid T = 1, M = m, \mathbf{X}, \mathbf{Z}) P(\mathbf{Z} \mid T = 1, \mathbf{X}) \\ - \mathbb{E}(Y \mid T = 0, M = m, \mathbf{X}, \mathbf{Z}) P(\mathbf{Z} \mid T = 0, \mathbf{X})\} P(\mathbf{X})$$

   - regression of Y on $T, M, \mathbf{X}, \mathbf{Z}$
   - model the distribution of $\mathbf{Z}$ given T and $\mathbf{X}$ $\rightsquigarrow$ difficult if $\mathbf{Z}$ is high-dimensional

2. No-interaction assumption

$$\mathbb{E}\{Y_i(t, m) - Y_i(t, m') \mid T_i = t, \mathbf{X}_i, \mathbf{Z}_i\} \\ = \mathbb{E}\{Y_i(t, m) - Y_i(t, m') \mid T_i = t, \mathbf{X}_i\}$$

   - causal effect of M on Y does not depend on $\mathbf{Z}$ given $T, \mathbf{X}$
   - structural nested mean models

Structural Nested Mean Models (Robins 1994. *Commun. Stat.*; see also Acharya et al. 2016 *Am. Political Sci. Rev.*)

1. Estimate the regression function

$$\mathbb{E}(Y_i \mid M_i, T_i, \mathbf{X}_i, \mathbf{Z}_i) \;=\; \alpha_0 + \alpha_1 T_i + \alpha_2 M_i + \boldsymbol{\alpha}_3^\top \mathbf{X}_i + \boldsymbol{\alpha}_4^\top \mathbf{Z}_i$$

   with no interaction between M and $\mathbf{Z}$ by the assumption

2. Compute the "blip-function"

$$\gamma(t, m, \mathbf{X}_i) \;=\; \mathbb{E}\{Y_i(t,m) - Y_i(t, m_0) \mid \mathbf{X}_i\} \;=\; \alpha_2(m - m_0)$$

   for any m representing the effect of $M = m$ (relative to m_0) on Y

3. Regress the adjusted outcome on T and $\mathbf{X}$

$$\mathbb{E}\{Y_i - \gamma(t, M_i, \mathbf{X}_i) \mid T_i, \mathbf{X}_i\} \;=\; \beta_0 + \underbrace{\beta_1}_{\bar{\xi}(m_0)} T_i + \boldsymbol{\beta}_2^\top \mathbf{X}_i$$

Linear Model and Natural Direct and Indirect Effects

- Linear structural equation model (LSEM):

$$
\begin{aligned}
Y_i &= \alpha_1 + \beta_1 T_i + \boldsymbol{\lambda}_1^\top \mathbf{X}_i + \epsilon_{1i} \\
M_i &= \alpha_2 + \beta_2 T_i + \boldsymbol{\lambda}_2^\top \mathbf{X}_i + \epsilon_{2i} \\
Y_i &= \alpha_3 + \beta_3 T_i + \gamma M_i + \boldsymbol{\lambda}_3^\top \mathbf{X}_i + \epsilon_{3i}
\end{aligned}
$$

  where the first equation is redundant

  1. Total effect is β_1
  2. Direct effect is β_3
  3. Indirect or mediation effect is $\beta_2\gamma = \beta_1 - \beta_3$
  4. Effect decomposition: $\beta_1 \;=\; \beta_3 + \beta_2\gamma$

- Baron and Kenny: distinction between moderation and mediation
- Moderated mediation:

$$
Y_i \;=\; \alpha_3 + \beta_3 T_i + \gamma M_i + \kappa T_i M_i + \boldsymbol{\lambda}_3^\top \mathbf{X}_i + \epsilon_{3i}
$$

  implying $\bar{\delta}(1) = \beta_2(\gamma + \kappa)$ and $\bar{\delta}(0) = \beta_2\gamma$

Estimation of Natural Direct and Indirect Effects

- Using the identification formula (NIE)

$$\bar{\delta}(t) = \sum_{M,\mathbf{X}} \mathbb{E}(Y \mid M, T = t, \mathbf{X}) \{P(M \mid T = 1, \mathbf{X}) - P(M \mid T = 0, \mathbf{X})\} \times P(\mathbf{X})$$

  1. predict M given each treatment value: $\{M_i(1), M_i(0)\}$
  2. predict Y by first setting $T_i = t$ and $M_i = M_i(0)$, and then $T_i = t$ and $M_i = M_i(1)$: $\{Y_i(t, M_i(0)), Y_i(t, M_i(1))\}$
  3. compute the average difference between two predicted outcomes

- NDE is similar but you can also estimate it by subtracting NIE from the total effect

$$\bar{\zeta}(t) = \sum_{M,\mathbf{X}} \{\mathbb{E}(Y \mid M, T = 1, \mathbf{X}) - \mathbb{E}(Y \mid M, T = 0, \mathbf{X})\} \times P(M \mid T = t, \mathbf{X}) P(\mathbf{X})$$