Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies

Kosuke Imai
Princeton University

The Fellowship of Woodrow Wilson Scholars Dinner Talk

February 17, 2014

Joint work with
Luke Keele   Dustin Tingley   Teppei Yamamoto

Identification of Causal Mechanisms

- Causal inference is a central goal of scientific research
- Scientists care about causal mechanisms, not just about causal effects

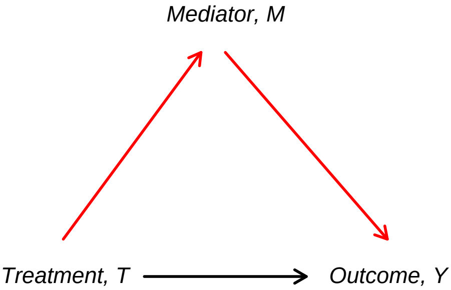
- Randomized experiments often only determine whether the treatment causes changes in the outcome
- Not how and why the treatment affects the outcome
- Common criticism of experiments and statistics:

black box view of causality

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

Causal Mediation Analysis

- Graphical representation



Mediator, M

Treatment, T

Outcome, Y

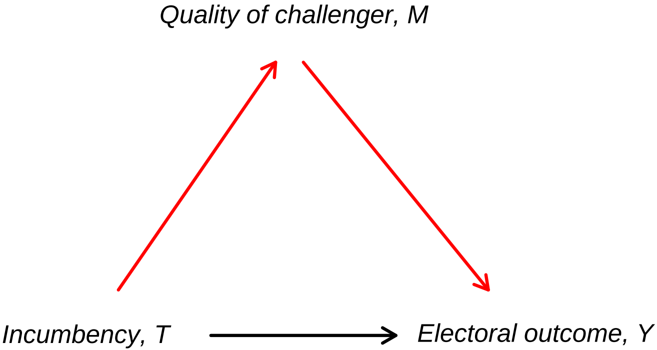
- Goal is to decompose total effect into direct and indirect effects
- Alternative approach: decompose the treatment into different components
- Causal mediation analysis as quantitative process tracing

Decomposition of Incumbency Advantage

- Incumbency effects: one of the most studied topics in American politics
- Consensus emerged in 1980s: incumbency advantage is positive and growing in magnitude

- New direction in 1990s: Where does incumbency advantage come from?
- Scare-off/quality effect (Cox and Katz): the ability of incumbents to deter high-quality challengers from entering the race
- Alternative causal mechanisms: name recognition, campaign spending, personal vote, television, etc.

Causal Mediation Analysis in Cox and Katz



Quality of challenger, M

Incumbency, T

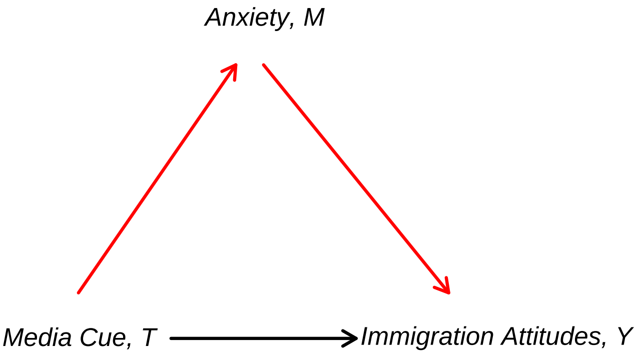
Electoral outcome, Y

- How much of incumbency advantage can be explained by scare-off/quality effect?
- How large is the mediation effect relative to the total effect?

Psychological Study of Media Effects

- Large literature on how media influences public opinion
- A media framing experiment of Brader *et al.*:
  1. (White) Subjects read a mock news story about immigration:
     - Treatment: Hispanic immigrant in the story
     - Control: European immigrant in the story
  2. Measure attitudinal and behavioral outcome variables:
     - Opinions about increasing or decrease immigration
     - Contact legislator about the issue
     - Send anti-immigration message to legislator
- Why is group-based media framing effective?: role of emotion
- Hypothesis: Hispanic immigrant increases anxiety, leading to greater opposition to immigration
- The primary goal is to examine how, not whether, media framing shapes public opinion

Causal Mediation Analysis in Brader *et al.*



- Does the media framing shape public opinion by making people anxious?
- An alternative causal mechanism: change in beliefs
- Can we identify mediation effects from randomized experiments?

The Standard Estimation Method

- Linear models for mediator and outcome:

$$\begin{aligned} Y_i &= \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \epsilon_{1i} \\ M_i &= \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{2i} \\ Y_i &= \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{3i} \end{aligned}$$

  where X_i is a set of pre-treatment or control variables
  1. Total effect (ATE) is β_1
  2. Direct effect is β_3
  3. Indirect or mediation effect is $\beta_2\gamma$
  4. Effect decomposition: $\beta_1 = \beta_3 + \beta_2\gamma$.

- Some motivating questions:
  1. What should we do when we have interaction or nonlinear terms?
  2. What about other models such as logit?
  3. In general, under what conditions can we interpret β_1 and $\beta_2\gamma$ as causal effects?
  4. What do we really mean by causal mediation effect anyway?

Potential Outcomes Framework of Causal Inference

- Observed data:
  - Binary treatment: $T_i \in \{0, 1\}$
  - Mediator: $M_i \in \mathcal{M}$
  - Outcome: $Y_i \in \mathcal{Y}$
  - Observed pre-treatment covariates: $X_i \in \mathcal{X}$
- Potential outcomes model (Neyman, Rubin):
  - Potential mediators: $M_i(t)$ where $M_i = M_i(T_i)$
  - Potential outcomes: $Y_i(t, m)$ where $Y_i = Y_i(T_i, M_i(T_i))$
- Total causal effect:

$$\tau_i \;\equiv\; Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

- **Fundamental problem of causal inference**: only one potential outcome can be observed for each *i*

Back to the Examples

- $M_i(1)$:
  1. Quality of her challenger if politician *i* is an incumbent
  2. Level of anxiety individual *i* would report if he reads the story with Hispanic immigrant

- $Y_i(1, M_i(1))$:
  1. Election outcome that would result if politician *i* is an incumbent and faces a challenger whose quality is $M_i(1)$
  2. Immigration attitude individual *i* would report if he reads the story with Hispanic immigrant and reports the anxiety level $M_i(1)$

- $M_i(0)$ and $Y_i(0, M_i(0))$ are the converse

Causal Mediation Effects

- Causal mediation (Indirect) effects:

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$

- Causal effect of the change in M_i on Y_i that would be induced by treatment
- Change the mediator from $M_i(0)$ to $M_i(1)$ while holding the treatment constant at t
- Represents the mechanism through M_i
- Zero treatment effect on mediator $\Longrightarrow$ Zero mediation effect

- Examples:
  1. Part of incumbency advantage that is due to the difference in challenger quality induced by incumbency status
  2. Difference in immigration attitudes that is due to the change in anxiety induced by the treatment news story

Total Effect $=$ Indirect Effect $+$ Direct Effect

- Direct effects:

$$\zeta_i(t) \;\equiv\; Y_i(1, M_i(t)) - Y_i(0, M_i(t))$$

- Causal effect of T_i on Y_i, holding mediator constant at its potential value that would realize when $T_i = t$
- Change the treatment from 0 to 1 while holding the mediator constant at $M_i(t)$
- Represents all mechanisms other than through M_i

- Total effect $=$ mediation (indirect) effect $+$ direct effect:

$$\tau_i \;=\; \delta_i(t) + \zeta_i(1-t) \;=\; \frac{1}{2}\{(\delta_i(0) + \zeta_i(0)) + (\delta_i(1) + \zeta_i(1))\}$$

Mechanisms, Manipulations, and Interactions

Mechanisms

- Indirect effects: $\delta_i(t) \;\equiv\; Y_i(t, M_i(1)) - Y_i(t, M_i(0))$
- Counterfactuals about treatment-induced mediator values

Manipulations

- Controlled direct effects: $\xi_i(t, m, m') \;\equiv\; Y_i(t, m) - Y_i(t, m')$
- Causal effect of directly manipulating the mediator under $T_i = t$

Interactions

- Interaction effects: $\xi(1, m, m') - \xi(0, m, m')$
- The extent to which controlled direct effects vary by the treatment

What Does the Observed Data Tell Us?

- Recall the Brader *et al.* experimental design:
  1. randomize T_i
  2. measure M_i and then Y_i

- Among observations with $T_i = t$, we observe $Y_i(t, M_i(t))$ but not $Y_i(t, M_i(1-t))$ unless $M_i(t) = M_i(1-t)$
- But we want to estimate

$$\delta_i(t) \;\equiv\; Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$

- For $t = 1$, we observe $Y_i(1, M_i(1))$ but not $Y_i(1, M_i(0))$
- Similarly, for $t = 0$, we observe $Y_i(0, M_i(0))$ but not $Y_i(0, M_i(1))$

- We have the identification problem $\Longrightarrow$ Need assumptions or better research designs

Counterfactuals in the Examples

1. Incumbency advantage:
   - An incumbent ($T_i = 1$) faces a challenger with quality $M_i(1)$
   - We observe the electoral outcome $Y_i = Y_i(1, M_i(1))$
   - We also want $Y_i(1, M_i(0))$ where $M_i(0)$ is the quality of challenger this incumbent politician would face if she is not an incumbent

2. Media framing effects:
   - A subject viewed the news story with Hispanic immigrant ($T_i = 1$)
   - For this person, $Y_i(1, M_i(1))$ is the observed immigration opinion
   - $Y_i(1, M_i(0))$ is his immigration opinion in the counterfactual world where he still views the story with Hispanic immigrant but his anxiety is at the same level as if he viewed the control news story

In both cases, we can't observe $Y_i(1, M_i(0))$ because $M_i(0)$ is not realized when $T_i = 1$

Project Goals (No Time Today to Cover the Details! ☹)

Provide a general framework for statistical analysis and research design strategies to understand causal mechanisms

1. Show that the sequential ignorability assumption is required to identify mechanisms even in experiments
2. Offer a flexible estimation strategy under this assumption
3. Introduce a sensitivity analysis to probe this assumption
4. Develop easy-to-use statistical software `mediation`
5. Propose research designs that relax sequential ignorability

Project References (click the article titles)

- **General:**
  - Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review*
- **Theory:**
  - Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*
- **Extensions:**
  - A General Approach to Causal Mediation Analysis. *Psychological Methods*
  - Experimental Designs for Identifying Causal Mechanisms. *Journal of the Royal Statistical Society, Series A (with discussions)*
  - Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments. *Political Analysis*
- **Software:**
  - mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*

The project website for papers and software:

`http://imai.princeton.edu/projects/mechanisms.html`

Email for questions and suggestions:

kimai@princeton.edu