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

Kosuke Imai
Princeton University

International Meeting in Experimental and Behavioral Social Sciences

April 11, 2014

Joint work with
Luke Keele   Dustin Tingley   Teppei Yamamoto

Project References

- **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*

Identification of Causal Mechanisms

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

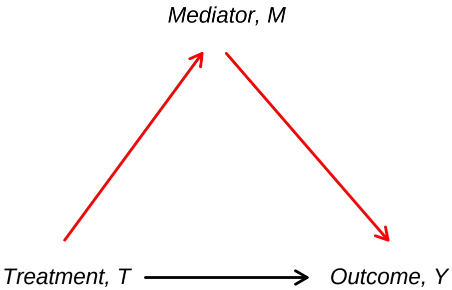
- Randomized experiments often only determine whether the treatment causes changes in the outcome
- Not how and why the treatment affects the outcome
- Common criticism of experiments and statistics:

black box view of causality

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

Causal Mediation Analysis

- Graphical representation



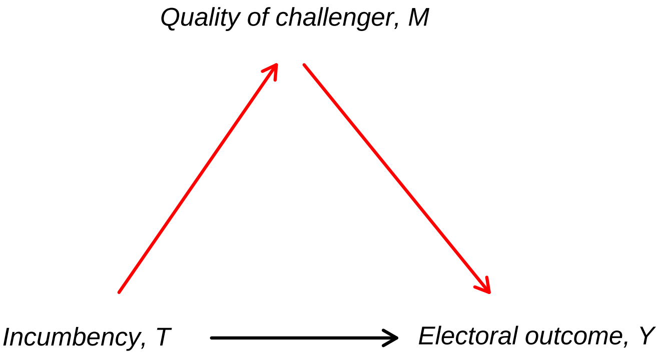
- Goal is to decompose total effect into direct and indirect effects
- Alternative approach: decompose the treatment into different components
- Causal mediation analysis as quantitative process tracing

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



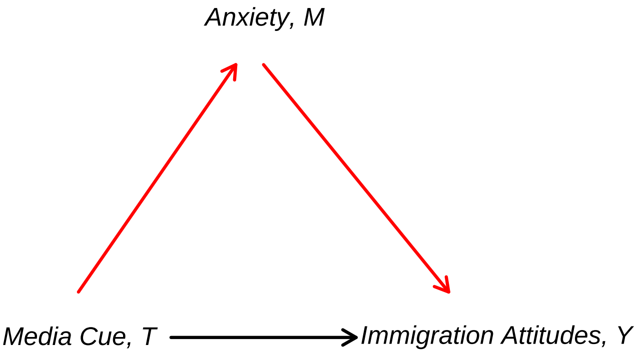
- How much of incumbency advantage can be explained by scare-off/quality effect?
- How large is the mediation effect relative to the total effect?

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

Sequential Ignorability Assumption

- Proposed identification assumption: Sequential Ignorability (SI)

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x, \qquad (1)$$

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x \qquad (2)$$

- In words,
  1. T_i is (as-if) randomized conditional on $X_i = x$
  2. $M_i(t)$ is (as-if) randomized conditional on $X_i = x$ and $T_i = t$

- Important limitations:
  1. In a standard experiment, (1) holds but (2) may not
  2. X_i needs to include all confounders
  3. X_i must be pre-treatment confounders $\Longrightarrow$ post-treatment confounder is not allowed
  4. Randomizing M_i via manipulation is not the same as assuming $M_i(t)$ is as-if randomized

Sequential Ignorability in the Standard Experiment

Back to Brader *et al.*:

- Treatment is randomized $\Longrightarrow$ (1) is satisfied
- But (2) may not hold:
  1. Pre-treatment confounder or X_i: state of residence
     those who live in AZ tend to have higher levels of perceived harm and be opposed to immigration
  2. Post-treatment confounder: alternative mechanism
     beliefs about the likely negative impact of immigration makes people anxious

- Pre-treatment confounders $\Longrightarrow$ measure and adjust for them
- Post-treatment confounders $\Longrightarrow$ adjusting is not sufficient

Nonparametric Identification

Under SI, both ACME and average direct effects are nonparametrically identified (can be consistently estimated without modeling assumption)

- ACME $\bar{\delta}(t)$

$$\int \int \mathbb{E}(Y_i \mid M_i, T_i = t, X_i) \{dP(M_i \mid T_i = 1, X_i) - dP(M_i \mid T_i = 0, X_i)\} \, dP(X_i)$$

- Average direct effects $\bar{\zeta}(t)$

$$\int \int \{\mathbb{E}(Y_i \mid M_i, T_i = 1, X_i) - \mathbb{E}(Y_i \mid M_i, T_i = 0, X_i)\} \, dP(M_i \mid T_i = t, X_i) \, dP(X_i)$$

Implies the general mediation formula under any statistical model

Beyond Sequential Ignorability

- Potential violations of sequential ignorability:
  1. unobserved pre-treatment confounder
  2. observed and unobserved post-treatment confounder

- Under the standard experimental design:
  - No-assumption bounds: even the sign of ACME is not identified
  - Sensitivity analysis: robustness of empirical findings to unobserved pre-treatment confounding
  - Statistical control: adjust for pre-treatment and post-treatment observed confounding

- Need for alternative experimental designs
- Possible when the mediator can be directly or indirectly manipulated
- New designs must preserve the ability to estimate the ACME under the SI assumption

Parallel Design

Randomly split sample

Experiment 1	**Experiment 2**
1) Randomize treatment	**1) Randomize treatment**
2) Measure mediator	**2) Randomize mediator**
3) Measure outcome	**3) Measure outcome**

- Must assume no direct effect of manipulation on outcome
- More informative than standard single experiment
- If we assume no T–M interaction, ACME is point identified

Why Do We Need No-Interaction Assumption?

- Numerical Example:

Prop.	$M_i(1)$	$M_i(0)$	$Y_i(t,1)$	$Y_i(t,0)$	$\delta_i(t)$
0.3	1	0	0	1	-1
0.3	0	0	1	0	0
0.1	0	1	0	1	1
0.3	1	1	1	0	0

- $\mathbb{E}(M_i(1) - M_i(0)) = \mathbb{E}(Y_i(t,1) - Y_i(t,0)) = 0.2$, but $\bar{\delta}(t) = -0.2$
- The Problem: Causal effect heterogeneity
  - T increases M only *on average*
  - M increases Y only *on average*
  - $T - M$ interaction: Many of those who have a positive effect of T on M have a negative effect of M on Y (first row)
- A solution: sensitivity analysis (see Imai and Yamamoto, 2013)
- Pitfall of “mechanism experiments” or “causal chain approach”

Example from Behavioral Neuroscience

Why study brain?: Social scientists' search for causal mechanisms underlying human behavior

- Psychologists, economists, and even political scientists

Question: What mechanism links low offers in an ultimatum game with "irrational" rejections?

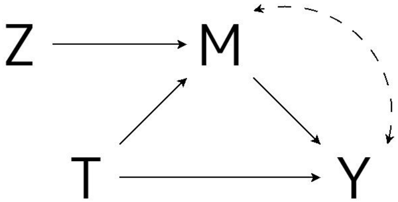
- A brain region known to be related to fairness becomes more active when unfair offer received (single experiment design)

Design solution: manipulate mechanisms with TMS

- Knoch et al. use TMS to manipulate — turn off — one of these regions, and then observes choices (parallel design)

Encouragement Design

- Direct manipulation of mediator is difficult in most situations
- Use an instrumental variable approach:



- Advantage: allows for unobserved confounder between *M* and *Y*
- Key Assumptions:
  1. *Z* is randomized or as-if random
  2. No direct effect of *Z* on *Y* (a.k.a. exclusion restriction)

Example: Social Norm Experiment on Property Taxes

- Lucia Del Carpio: “Are Neighbors Cheating?”
- Treatment: informing compliance rate of neighbors
- Most people underestimate compliance rate
- Outcome: compliance rate obtained from administrative records
- Large positive effect on compliance rate $\approx$ 20 percentage points

- Mechanisms:
  1. $M =$ beliefs about enforcement (measured)
  2. social norm (not measured; direct effect)
- Instrument: $Z =$ informing enforcement rate
- Assumption: Z affects Y only through M

- Results:
  - Average direct effect is estimated to be large
  - The author interprets this effect as the effect of social norm

Crossover Design

- Recall ACME can be identified if we observe $Y_i(t', M_i(t))$
- Get $M_i(t)$, then switch T_i to t' while holding $M_i = M_i(t)$

- Crossover design:
  1. Round 1: Conduct a standard experiment
  2. Round 2: Change the treatment to the opposite status but fix the mediator to the value observed in the first round
- Very powerful – identifies mediation effects for each subject
- Must assume no carryover effect: Round 1 must not affect Round 2
- Can be made plausible by design

Example: Labor Market Discrimination

EXAMPLE Bertrand & Mullainathan (2004, AER)

- Treatment: Black vs. White names on CVs
- Mediator: Perceived qualifications of applicants
- Outcome: Callback from employers

- Quantity of interest: Direct effects of (perceived) race
- Would Jamal get a callback if his name were Greg but his qualifications stayed the same?

- Round 1: Send Jamal's actual CV and record the outcome
- Round 2: Send his CV as Greg and record the outcome

- Assumption: their different names do not change the perceived qualifications of applicants
- Under this assumption, the direct effect can be interpreted as blunt racial discrimination

Cross-over Design in Observational Studies

Experimental design as a template for observational studies

EXAMPLE Back to incumbency advantage

- Use of cross-over design (Levitt and Wolfram)
  1. 1st Round: two non-incumbents in an open seat
  2. 2nd Round: same candidates with one being an incumbent
- Assume challenger quality (mediator) stays the same
- Estimation of direct effect is possible

- Redistricting as natural experiments (Ansolabehere et al.)
  1. 1st Round: incumbent in the old part of the district
  2. 2nd Round: incumbent in the new part of the district
- Challenger quality is the same but treatment is different
- Estimation of direct effect is possible

Concluding Remarks

- Even in a randomized experiment, a strong assumption is needed to identify causal mechanisms
- However, progress can be made toward this fundamental goal of scientific research with modern statistical tools
- A general, flexible estimation method is available once we assume sequential ignorability
- Sequential ignorability can be probed via sensitivity analysis
- More credible inferences are possible using clever experimental designs
- Insights from new experimental designs can be directly applied when designing observational studies
- Multiple mediators require additional care when they are causally dependent

The project website for papers and software:

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

Email for questions and suggestions:

kimai@princeton.edu