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

Kosuke Imai

Princeton University

Joint work with
Keele (Ohio State), Tingley (Harvard), Yamamoto (Princeton)

January 19, 2011

Harris School of Public Policy
The University of Chicago

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?

Goals of the Talk

Present a general framework for statistical design and analysis of 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. Propose a sensitivity analysis to probe this assumption
4. Propose new experimental designs that do not rely on sequential ignorability
5. Extend these methods to observational studies

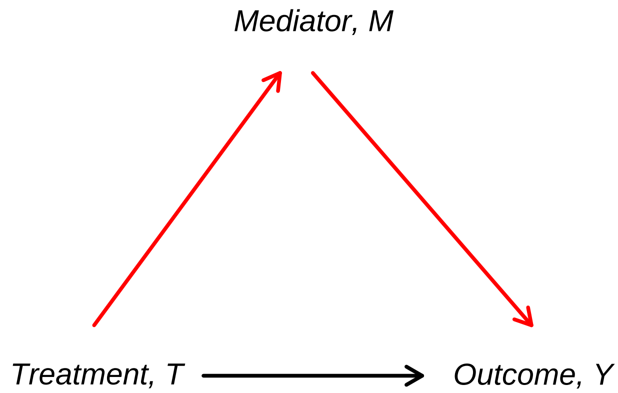
Project Reference

- Project Website:

  http://imai.princeton.edu/projects/mechanisms.html
- Papers:
  - "Unpacking the Black Box: Learning about Causal Mechanisms from Experimental and Observational Studies."
  - "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science*
  - "A General Approach to Causal Mediation Analysis." *Psychological Methods*
  - "Experimental Identification of Causal Mechanisms."
  - "Causal Mediation Analysis Using R." *Advances in Social Science Research Using R*
- Software: R package `mediation` implements all methods

What Is a Causal Mechanism?

- Mechanisms as alternative causal pathways
- Causal mediation analysis



- Quantities of interest: Direct and indirect effects
- Fast growing methodological literature

Two American Politics Examples

1. Media priming
   - Media cues referencing ethnic groups effectively affect attitudes towards immigration policy
   - Brader, Valentino, and Suhay (2008, AJPS): Anxiety transmits the effect of cues on attitudes
   - Experimental study with randomized treatment
2. Incumbency advantage
   1. Imcumbency advantage has been positive and growing
   2. Cox and Katz (1996, AJPS): incumbents deter high-quality challengers from entering the race
   3. Observational study with non-random treatment

Potential Outcomes Framework

Framework: Potential outcomes model of causal inference

- 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 mediators: $M_i(t)$, where $M_i = M_i(T_i)$ observed
- Potential outcomes: $Y_i(t, m)$, where $Y_i = Y_i(T_i, M_i(T_i))$ observed
- In a standard experiment, **only one potential outcome** can be observed for each i

Causal Mediation Effects

- Total causal effect:

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

- 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

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:

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

What Does the Observed Data Tell Us?

- Quantity of Interest: Average causal mediation effects

$$\bar{\delta}(t) \equiv \mathbb{E}(\delta_i(t)) = \mathbb{E}\{Y_i(t, M_i(1)) - Y_i(t, M_i(0))\}$$

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

- Problem: $Y_i(t, M_i(t))$ is observed but $Y_i(t, M_i(t'))$ can never be observed
- We have an identification problem

$\Longrightarrow$ Need additional assumptions to make progress

Identification under Sequential Ignorability

- Proposed identification assumption: Sequential Ignorability

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

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

- (1) is guaranteed to hold in a standard experiment
- (2) does **not** hold unless X_i includes all confounders

Theorem: Under sequential ignorability, ACME and average direct effects are nonparametrically identified
($=$ consistently estimated from observed data)

Nonparametric Identification

Theorem: Under SI, both ACME and average direct effects are given by,

- 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 $\bar{\zeta}(t)$

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

Traditional Estimation Method

- Linear structural equation model (LSEM):

$$\begin{aligned} M_i &= \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{i2}, \\ Y_i &= \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{i3}. \end{aligned}$$

  together implying

$$Y_i = \alpha_1 + \beta_1 T_i + \epsilon_{i1}$$

- Fit two least squares regressions separately
- Use product of coefficients ($\hat{\beta}_2 \hat{\gamma}$) to estimate ACME
- Use asymptotic variance to test significance (Sobel test)

- Under SI and the no-interaction assumption ($\bar{\delta}(1) \neq \bar{\delta}(0)$), $\hat{\beta}_2 \hat{\gamma}$ consistently estimates ACME
- Can be extended to LSEM with interaction terms

- Problem: Only valid for the simplest LSEM

Proposed General Estimation Algorithm

1. Model outcome and mediator
   - Outcome model: $p(Y_i \mid T_i, M_i, X_i)$
   - Mediator model: $p(M_i \mid T_i, X_i)$
   - These models can be of any form (linear or nonlinear, semi- or nonparametric, with or without interactions)

2. Predict mediator for both treatment values ($M_i(1)$, $M_i(0)$)
3. Predict outcome by first setting $T_i = 1$ and $M_i = M_i(0)$, and then $T_i = 1$ and $M_i = M_i(1)$
4. Compute the average difference between two outcomes to obtain a consistent estimate of ACME

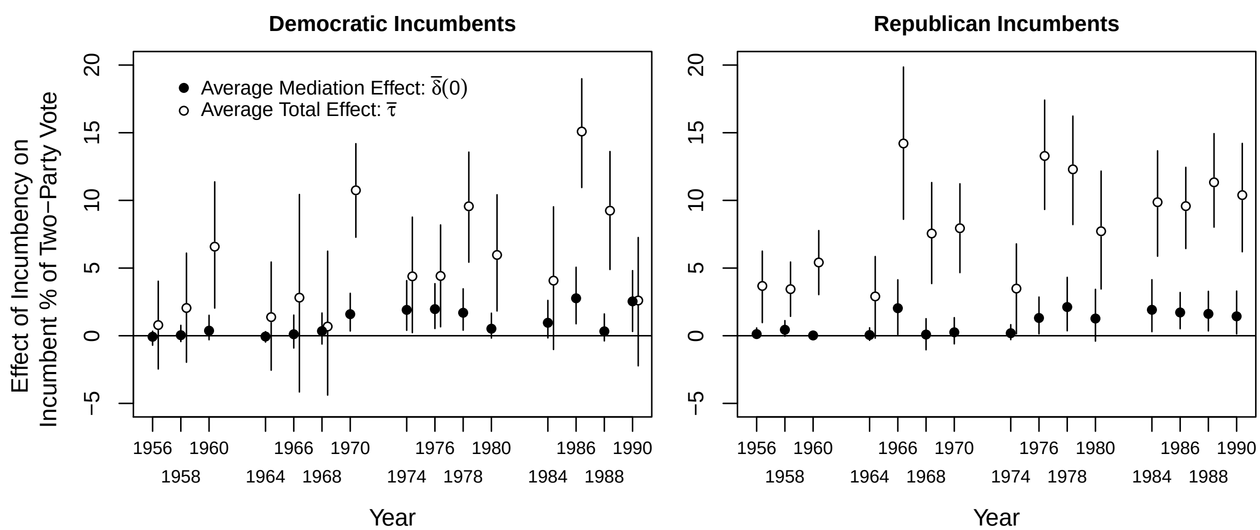
5. Monte Carlo simulation or bootstrap to estimate uncertainty

Reanalysis of Brader et al. (2008)

- ACME $=$ Average difference in immigration attitudes due to the change in anxiety induced by the media cue treatment
- Sequential ignorability $=$ No unobserved covariate affecting both anxiety and immigration attitudes
- Original method: Product of coefficients with the Sobel test — Valid only when both models are linear w/o T–M interaction (which they are not)
- Our method: Calculate ACME using our general algorithm

	Product of Coefficients	Average Causal Mediation Effect
Decrease Immigration	.399 [0.066, .732]	.089 [0.023, .178]
Request Anti-Immigration Info	.295 [0.023, 0.567]	.049 [0.007, 0.121]
Send Anti-Immigration Message	.303 [0.046, .561]	.105 [0.021, 0.191]

Reanalysis of Cox and Katz (1996)



- Original findings:
  1. Incumbency advantage has increased over time
  2. This increase is attributed to increase in scare-off/quality effect
- Our findings: Increasing incumbency advantage may be attributable to mechanisms other than the scare-off/quality effect

Need for Sensitivity Analysis

- Standard experiments require sequential ignorability to identify mechanisms
- The sequential ignorability assumption is often too strong
- Need to assess the robustness of findings via sensitivity analysis
- Question: How large a departure from the key assumption must occur for the conclusions to no longer hold?
- Parametric sensitivity analysis by assuming

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

but not

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

- Possible existence of unobserved *pre-treatment* confounder

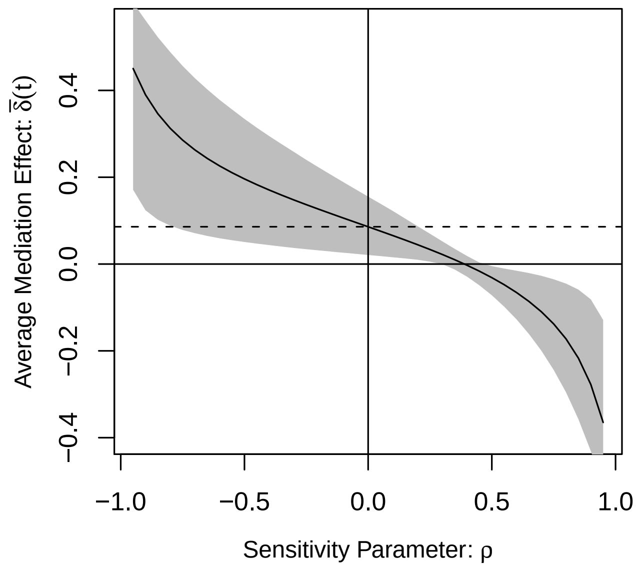
Parametric Sensitivity Analysis

- Sensitivity parameter: $\rho \equiv \text{Corr}(\epsilon_{i2}, \epsilon_{i3})$
- Sequential ignorability implies $\rho = 0$
- Set ρ to different values and see how ACME changes
- Interpreting ρ: how small is too small?
- An unobserved (pre-treatment) confounder formulation:

$$\epsilon_{i2} = \lambda_2 U_i + \epsilon'_{i2} \quad \text{and} \quad \epsilon_{i3} = \lambda_3 U_i + \epsilon'_{i3}$$

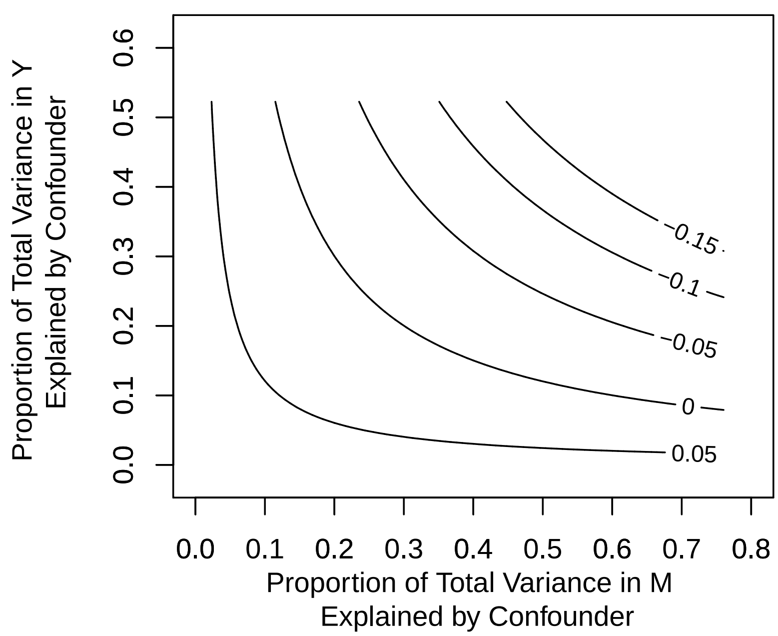
- How much does U_i have to explain for our results to go away?

Sensitivity Analysis of Brader *et al.* (2008) I



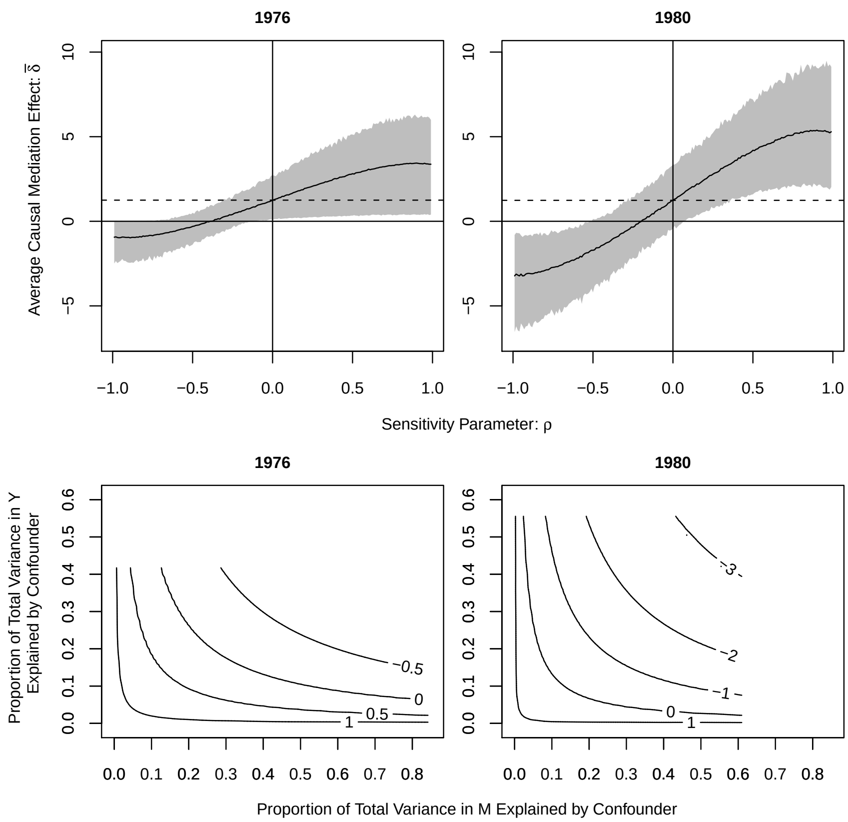
- ACME $>$ 0 as long as the error correlation is less than 0.39 (0.30 with 95% CI)

Sensitivity Analysis of Brader *et al.* (2008) II



- An unobserved confounder can account for up to 26.5% of the variation in both Y_i and M_i before ACME becomes zero

Sensitivity Analysis of Cox and Katz (1996)



Beyond Sequential Ignorability

- Without sequential ignorability, standard experimental design lacks identification power
- Even the sign of ACME is not identified
- Need to develop alternative experimental designs for more credible inference
- Possible when the mediator can be directly or indirectly manipulated

Parallel Design

Randomly split sample

Experiment 1

1) Randomize treatment

2) Measure mediator

3) Measure outcome

Experiment 2

1) Randomize treatment

2) Randomize mediator

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

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)

Limitations

- Difference between manipulation and mechanism

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

- Here, $\mathbb{E}(M_i(1) - M_i(0)) = \mathbb{E}(Y_i(t,1) - Y_i(t,0)) = 0.2$, but $\bar{\delta}(t) = -0.2$

- Limitations:
  - Direct manipulation of the mediator is often impossible
  - Even if possible, manipulation can directly affect outcome
- Need to allow for subtle and indirect manipulations

Encouragement Design

- Randomly encourage subjects to take particular values of the mediator M_i
- Standard instrumental variable assumptions (Angrist et al.)

Use a 2×3 factorial design:

1. Randomly assign T_i
2. Also randomly decide whether to **positively encourage**, **negatively encourage**, or do nothing
3. Measure mediator and outcome

- Informative inference about the “complier” ACME
- Reduces to the parallel design if encouragement is perfect
- Possible application to the immigration experiment: Use autobiographical writing tasks to encourage anxiety

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 from Labor Economics

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

- No carryover effect: send two CVs to randomly selected, different potential employers
- Assumption: perceived qualifications don't depend on applicant's race

Designing Observational Studies

- Key difference between experimental and observational studies: treatment assignment
- Sequential ignorability:
  1. Ignorability of treatment given covariates
  2. Ignorability of mediator given treatment and covariates
- Both (1) and (2) are suspect in observational studies

- Statistical control: matching, regressions, etc.
- Search for quasi-randomized treatments: “natural” experiments

- How can we design observational studies?
- Experiments can serve as templates for observational studies

Example from Imcumbency 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