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

Kosuke Imai
Department of Politics
Center for Statistics and Machine Learning
Princeton University

Keynote Talk
The Quality Registry Research Conference
Stockholm, Sweden

May 24, 2016

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 $\rightsquigarrow$ 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:

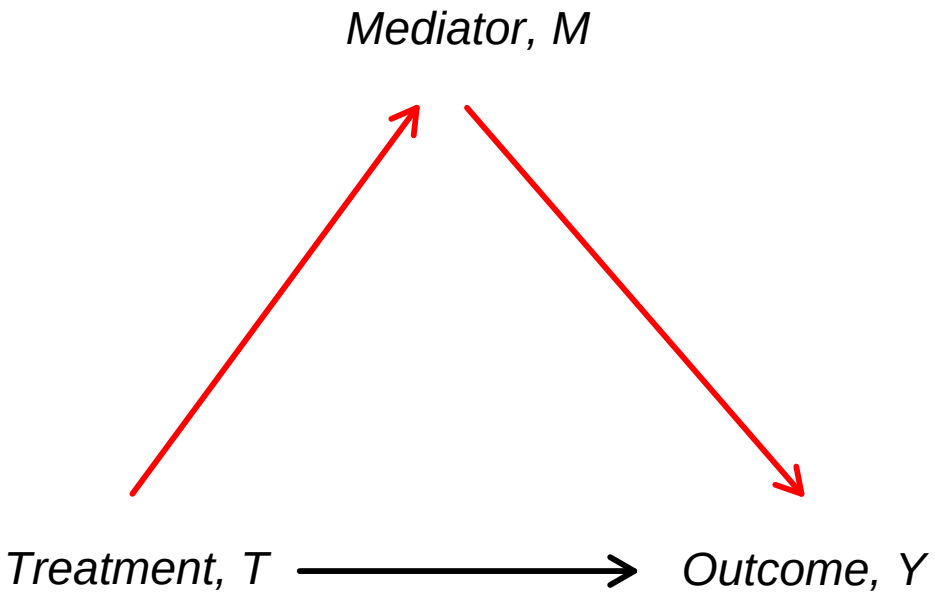
black box view of causality

- Qualitative research $\rightsquigarrow$ process tracing

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

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
- How large is the mediation effect relative to the total effect?

Mexican Universal Health Insurance Program

- Seguro Popular (2003): cover all 50M uninsured Mexicans
- Matched-pair cluster randomized design
- Papers in Lancet and Statistical Science (2009)
- Treatment T:
  - building hospitals and clinics
  - encouragement to sign up for SP
- Post-treatment measures:
  - financial protection
  - healthcare utilization
  - health
- Mediation analysis:
  - M: reduction in catastrophic expenditure
  - Y: health outcome

Decomposition of Incumbency Advantage

- Incumbency effects: one of the most studied topics
- 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: the ability of incumbents to deter high-quality challengers from entering the race
- Alternative causal mechanisms: name recognition, campaign spending, personal vote, television, etc.

- Mediation analysis:
  - T: incumbency status
  - M: quality of challenger
  - Y: election outcome

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. Level of catastrophic health expenditure for an individual i
  2. Quality of challenger if politician i is an incumbent

- $Y_i(1, M_i(1))$:
  1. Health outcome that would result if individual i pays catastrophic health expenditure $M_i(1)$
  2. Election outcome that would result if politician i is an incumbent and faces a challenger whose quality is $M_i(1)$

- $M_i(0)$ 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 health effects that are due to the reduction in the level of catastrophic expenditure
  2. Part of incumbency advantage that is due to the difference in challenger quality induced by incumbency status

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. Health insurance evaluation:
   - An individual lives in the treatment community ($T_i = 1$)
   - For this person, $Y_i(1, M_i(1))$ is the observed health outcome
   - $Y_i(1, M_i(0))$ is his health outcome in the counterfactual world where he still lives in the treatment village but his catastrophic expenditure is at the same level as it would be if his village did not receive the treatment

2. 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

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 Seguro Popular:

- Treatment is randomized $\Longrightarrow$ (1) is satisfied
- But (2) may not hold:
  1. Pre-treatment confounder or X_i: health predisposition
     people with poor health are more likely to pay catastrophic health expenditure and have poor health in the future
  2. Post-treatment confounder: alternative mechanism
     Seguro Popular increases the use of preventive care, which in turn reduces catastrophic expenditure and improves future health outcome

- 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

Traditional Estimation Methods: LSEM

- Linear structural equation model (LSEM):

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

- 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

Popular Baron-Kenny Procedure

- The procedure:
  1. Regress Y on T and show a significant relationship
  2. Regress M on T and show a significant relationship
  3. Regress Y on M and T, and show a significant relationship between Y and M

- The problems:
  1. First step can lead to false negatives especially if indirect and direct effects in opposite directions
  2. The procedure only anticipates simplest linear models
  3. Don't do star-gazing. Report quantities of interest

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 or bootstrap to estimate uncertainty

Example: Binary Mediator and Outcome

- Two logistic regression models:

$$\begin{aligned} \Pr(M_i = 1 \mid T_i, X_i) &= \text{logit}^{-1}(\alpha_2 + \beta_2 T_i + \xi_2^\top X_i) \\ \Pr(Y_i = 1 \mid T_i, M_i, X_i) &= \text{logit}^{-1}(\alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i) \end{aligned}$$

- Can't multiply β_2 by γ
- Difference of coefficients $\beta_1 - \beta_3$ doesn't work either

$$\Pr(Y_i = 1 \mid T_i, X_i) = \text{logit}^{-1}(\alpha_1 + \beta_1 T_i + \xi_1^\top X_i)$$

- Can use our algorithm (example: $\mathbb{E}\{Y_i(1, M_i(0))\}$)
  1. Predict $M_i(0)$ given $T_i = 0$ using the first model
  2. Compute $\Pr(Y_i(1, M_i(0)) = 1 \mid T_i = 1, M_i = \widehat{M}_i(0), X_i)$ using the second model

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
- Another sensitivity parameter: R^2

- **Result:**

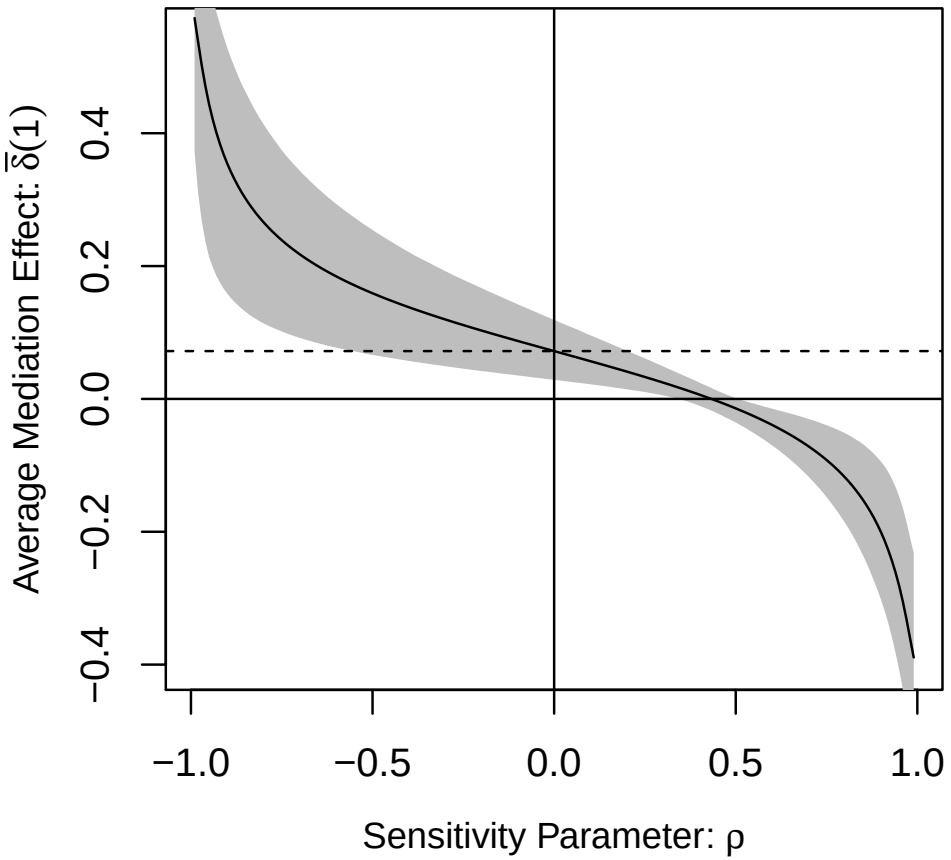
$$\bar{\delta}(0) = \bar{\delta}(1) = \frac{\beta_2 \sigma_1}{\sigma_2} \left\{ \tilde{\rho} - \rho \sqrt{(1 - \tilde{\rho}^2)/(1 - \rho^2)} \right\},$$

where $\sigma_j^2 \equiv \text{var}(\epsilon_{ij})$ for $j = 1, 2$ and $\tilde{\rho} \equiv \text{Corr}(\epsilon_{i1}, \epsilon_{i2})$.

- When do my results go away completely?
- $\bar{\delta}(t) = 0$ if and only if $\rho = \tilde{\rho}$
- Easy to estimate from the regression of Y_i on T_i:

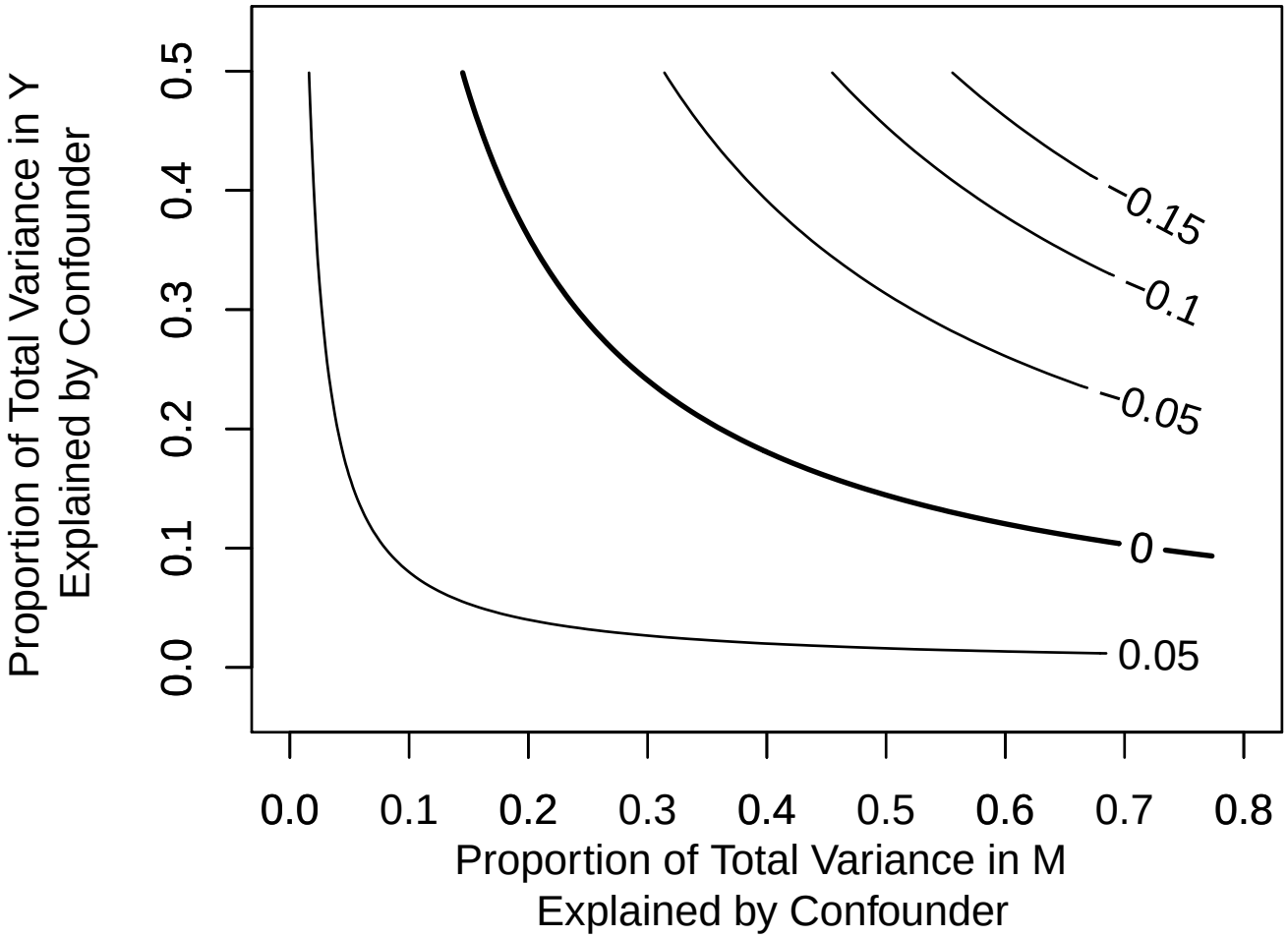
$$Y_i = \alpha_1 + \beta_1 T_i + \epsilon_{i1}$$

Sensitivity Analysis Using ρ



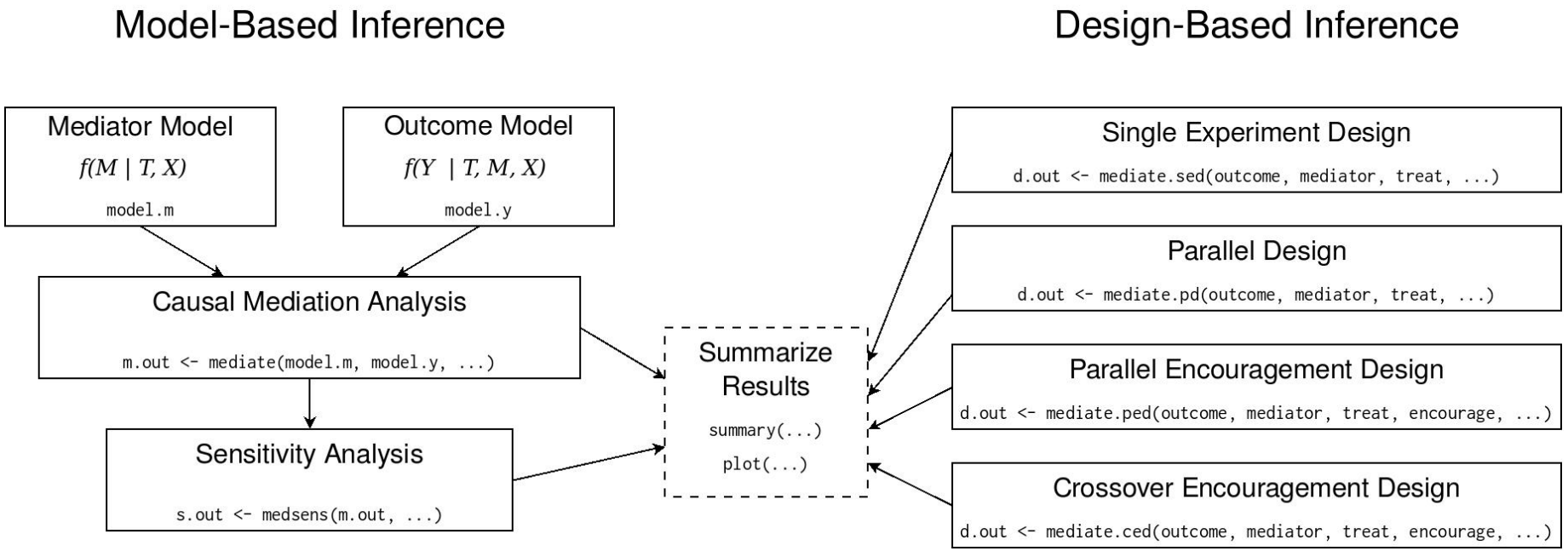
- ACME $>$ 0 as long as the error correlation is less than 0.39 (0.30 with 95% CI)

Sensitivity Analysis Using $\tilde{R}^2_M$ and $\tilde{R}^2_Y$



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

Open-Source Software “Mediation”



Implementation Examples

1. Fit models for the mediator and outcome variable and store these models

```
> m <- lm(Mediator ~ Treat + X)
> y <- lm(Y ~ Treat + Mediator + X)
```

2. Mediation analysis: Feed model objects into the `mediate()` function. Call a summary of results

```
> m.out<-mediate(m, y, treat = "Treat",
                 mediator = "Mediator")
> summary(m.out)
```

3. Sensitivity analysis: Feed the output into the `medsens()` function. Summarize and plot

```
> s.out <- medsens(m.out)
> summary(s.out)
> plot(s.out, "rho")
> plot(s.out, "R2")
```

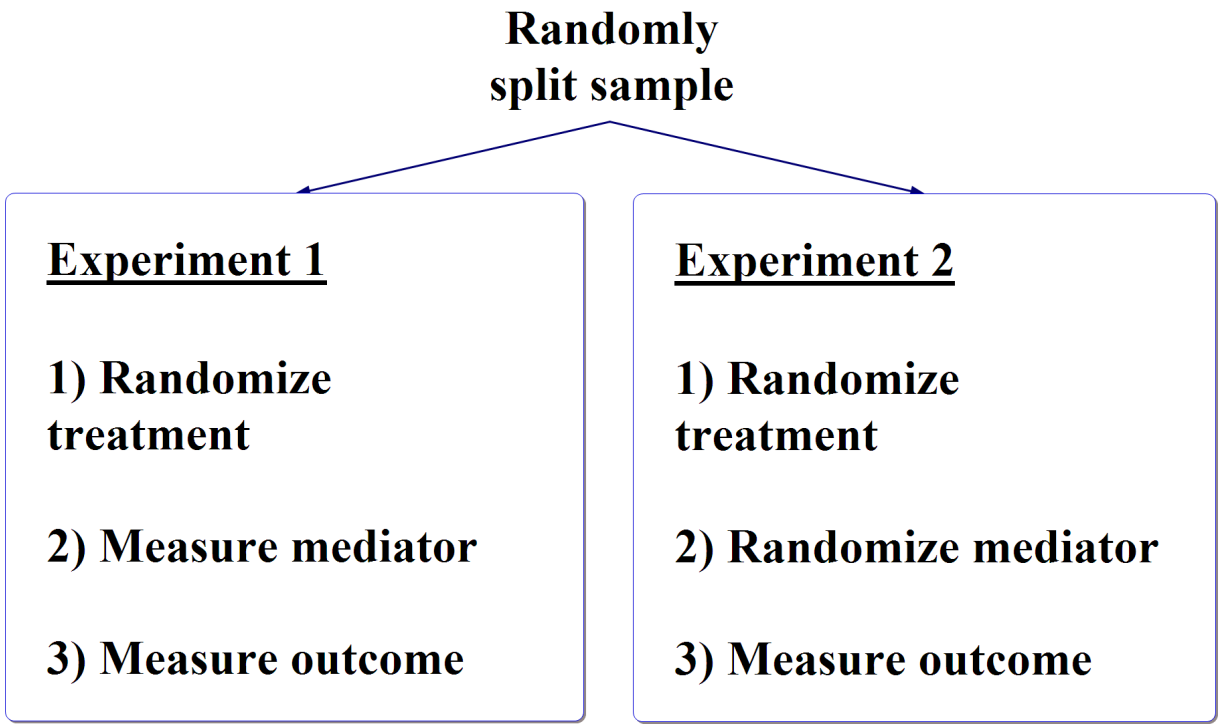
Beyond Sequential Ignorability

- 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



- 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?

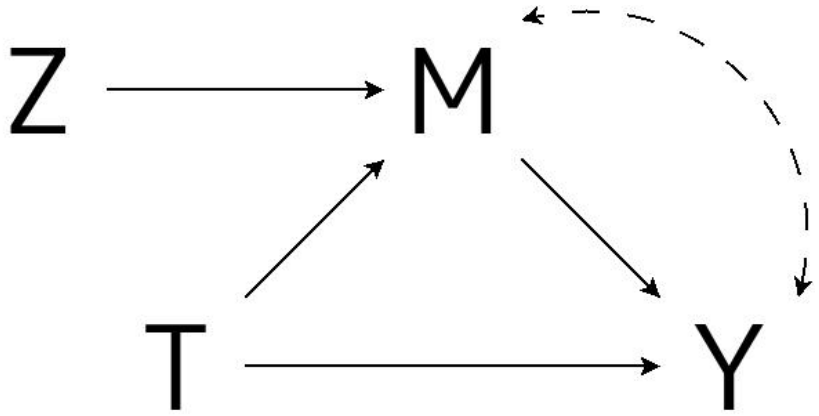
- A brain region known to be related to fairness becomes more active when unfair offer received (single experiment design)

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

Do experiments have any value without mediation?

- Yes, but it is crucial to understand mechanisms:
  - scientists want to test theories which are about mechanisms
  - policy makers want to devise better policies
  - understanding of mechanisms $\rightsquigarrow$ external validity

- Two ways to address the question, "why does a treatment work?"
  1. mediation $\rightsquigarrow$ causal process
  2. interaction $\rightsquigarrow$ causal components

What do you think about mechanism experiments?

- "mechanism experiments" (Ludwig, Kling, and Mullainathan, 2011)
- "causal chain approach" (Spencer, Zanna, and Fong, 2005)
  1. Randomize T to identify its effect on Y and its effect on M
  2. Randomize M to identify its effect on Y
- This is certainly a progress towards understanding mechanisms

- Two issues with this approach (Imai, Tingley, and Yamamoto, JRSSA, 2013):
  1. Effects of direct manipulation of M may differ from those of "natural" change in M induced by T
  2. Effect heterogeneity: even if the average effect of T on M and that of M on Y are both positive, the average mediation effect of T on Y can be negative

How sensitive do the results of sensitivity analysis have to be before doubting mediation analysis?

- What sensitivity analysis provides: the amount of hidden bias that makes one's mediational results go away
- Traditional tests: sampling uncertainty of one's mediational effects that are assumed to be identifiable with the infinite amount of data
- Can a scientific community agree on the required degree of sensitivity? $\rightsquigarrow$ maybe not
- Rosenbaum's example:
  1. Effect of smoking on cancer: $\Gamma = 6$
  2. Effect of coffee on myocardial infarction: $\Gamma = 1.3$
- Need to accumulate sensitivity analysis results
- Need to look for confounders that reduce sensitivity

Other Questions

1. Why can't we just show those who have the large effects of T on M also exhibit the large effects of M on Y?
   - Yes, but those effects must be identified
   - Reducing heterogeneity helps the identification of mediation effects

2. Is mediation analysis uninformative because it can hardly be definitive?
   - No. Almost no scientific study can be definitive.
   - But mediation is about purely counterfactual quantities

3. What researchers can do to maximize the plausibility of sequential ignorability?
   - Better design with clever manipulation of mediators
   - Importance of sensitivity analysis

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*

The project website for papers and software:

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

Email for questions and suggestions:

kimai@princeton.edu