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

Kosuke Imai

Princeton University

December 6, 2012
University of Wisconsin, Madison
Center for Education Research

Project References

This talk is based on the following papers and software:

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

All of these and other materials are available at

http://imai.princeton.edu/projects/mediation

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?

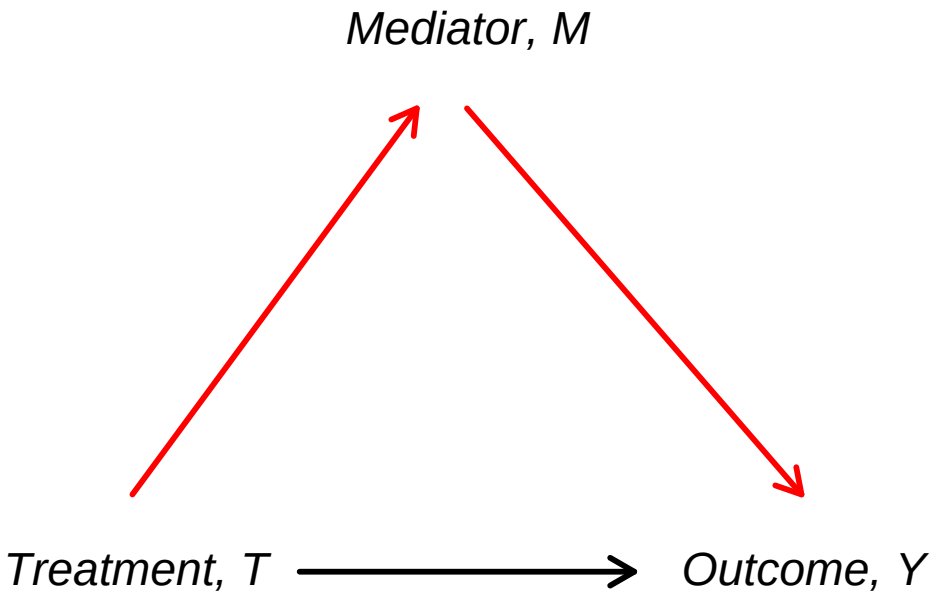
Goals of the Talk

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. Illustrate how to use the open-source software **mediation**
5. Propose new experimental designs that do not rely on sequential ignorability
6. Cover both experiments and observational studies under the same principle

Causal Mediation Analysis

- Causal mechanisms as causal pathways
- Graphical representation



- Goal is to decompose total effect into direct and indirect effects
- Example in education research:
  - Treatment: New curriculum
  - Mediator: Teachers' efforts
  - Outcome: Students' test score

Standard Estimation Methods

- Standard Linear Structural Equation Models:

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

- Total effect (ATE) is β_1
- Direct effect is β_3
- Indirect or mediation effect is: $\beta_2\gamma$ or $\beta_1 - \beta_3$
- Effect decomposition: $\beta_1 = \beta_3 + \beta_2\gamma$

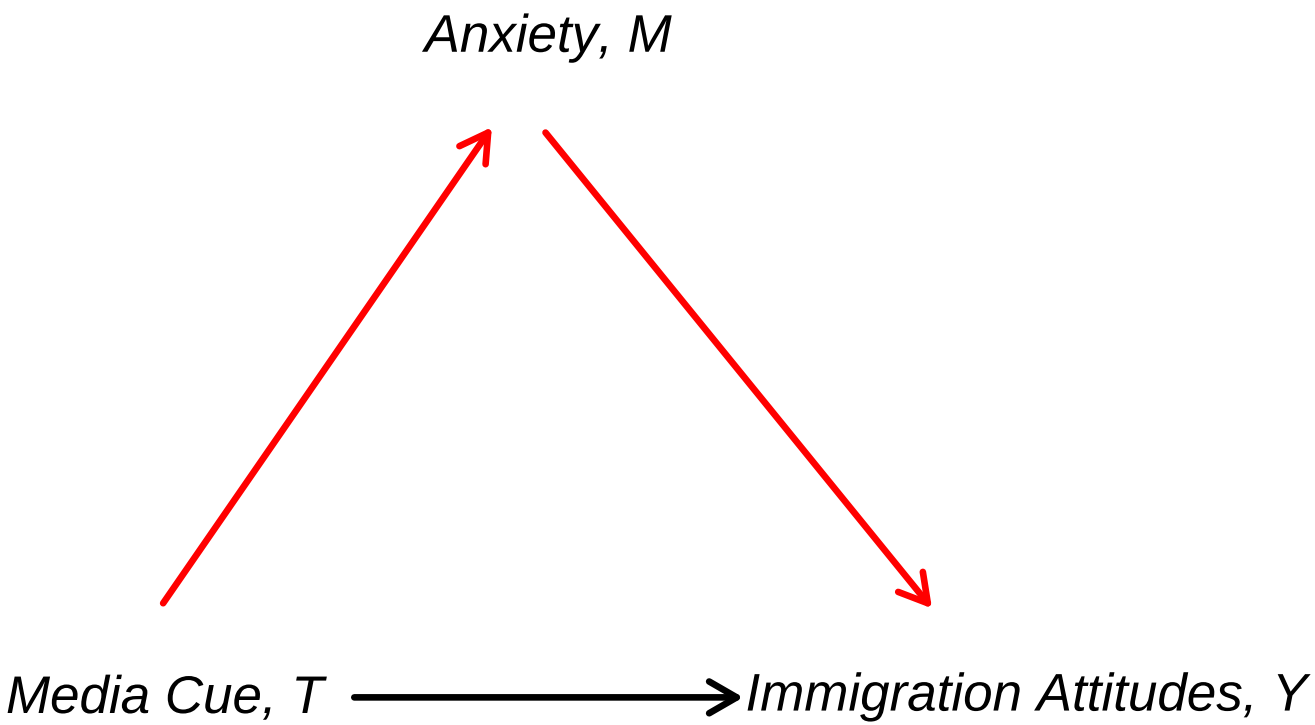
- But what are we assuming when we do mediation analysis?
- How sensitive our results are to the violation of this assumption?
- What should we do if our model is not a simple linear regression?
- Are there research designs that require weaker assumptions?

Media Cues and Immigration Attitudes

Brader et al. experiment:

- (White) Subjects read a mock news story about immigration
- Treatment: Hispanic immigrant in the story
- Control: European immigrant in the story
- Attitudinal and behavioral outcome variables:
  - Opinions about increasing or decrease immigration
  - Contact legislator about the issue
  - Send anti-immigration message to legislator
- Hypothesis: Hispanic immigrant increases anxiety, leading to greater opposition to immigration

Causal Mediation Analysis in Brader et al.



What's the effect of the news story that works by making people anxious?

Potential Outcomes Framework 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

- Typically, only one potential outcome (mediator) is observed

Media Cue Study with this Notation

- $T_i = 1$: Adam "i" was (randomly) assigned to the treatment group

- $M_i(1) = M_i$: observed level of anxiety reported by Adam
- $M_i(0)$: counterfactual level of Adam's anxiety if assigned to the control condition

- $Y_i(1, M_i(1)) = Y_i$: observed immigration attitude reported by Adam
- Suppose Adam got anxious: $M_i = M_i(1) = 1$
  - $Y_i(1, 1)$: observed outcome
  - $Y_i(1, 0)$, $Y_i(0, 1)$, $Y_i(0, 0)$: counterfactual outcomes

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 causal 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 causal mechanisms other than through M_i

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

$$\begin{aligned}
\tau_i &= \underbrace{Y_i(1, M_i(1)) - Y_i(1, M_i(0))}_{\delta_i(1)} + \underbrace{Y_i(1, M_i(0)) - Y_i(0, M_i(0))}_{\zeta_i(0)} \\
&= \underbrace{Y_i(1, M_i(1)) - Y_i(0, M_i(1))}_{\zeta_i(0)} + \underbrace{Y_i(0, M_i(1)) - Y_i(0, M_i(0))}_{\delta_i(1)}
\end{aligned}$$

Back to the Media Cue Study

- Mediation effect: Treatment effect on immigration attitudes that is due to the change in anxiety induced by the treatment news story

- Scenario I: Treatment makes Adam anxious $(M_i(1), M_i(0)) = (1, 0)$
  - $\delta_i(1) = Y_i(1,1) - Y_i(1,0)$
  - $\delta_i(0) = Y_i(0,1) - Y_i(0,0)$
  - $\zeta_i(1) = Y_i(1,1) - Y_i(0,1)$
  - $\zeta_i(0) = Y_i(1,0) - Y_i(0,0)$

- Scenario II: Adam is always anxious $(M_i(1), M_i(0)) = (1, 1)$
  - $\delta_i(1) = Y_i(1,1) - Y_i(1,1) = 0$
  - $\delta_i(0) = Y_i(0,1) - Y_i(0,1) = 0$
  - $\zeta_i(1) = Y_i(1,1) - Y_i(0,1)$
  - $\zeta_i(0) = Y_i(0,1) - Y_i(0,1)$
- Zero treatment effect on mediator $\Longrightarrow$ zero mediation effect

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: $\xi(1, m, m') - \xi(0, m, m') \neq 0$
- Causal effect of mediator on outcome depends on treatment
- Doesn't imply the existence of a mechanism

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
- Adam received the treatment:
  - $Y_i(1, M_i(1))$ is observed
  - $Y_i(0, M_i(0))$ is not observed but is potentially observable
  - $Y_i(1, M_i(0))$ and $Y_i(0, M_i(1))$ are pure counterfactuals
- We have an identification problem!

$\Longrightarrow$ Need additional assumptions to make progress

Sequential Ignorability Assumption

- Proposed identification assumption: Sequential Ignorability (SI)

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

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

- (1) is guaranteed to hold in a standard experiment
- (2) does **not** hold unless X_i includes all confounders
- X_i has to be pre-treatment confounders
- Post-treatment confounder is not allowed $\Longrightarrow$ alternative mediators are allowed but they cannot be causally related (more on this later)

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 model

Traditional Estimation Method

- Linear structural equation model (LSEM):

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

- 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 or bootstrapping 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 $\Pr(Y_i(1, M_i(0)) = 1 \mid X_i)$ given $T_i = 1$ and $\widehat{M}_i(0)$ using the second model

Sensitivity Analysis

- Standard experiments require SI to identify mechanisms
- This assumption is often too strong

- Need to assess the robustness of findings via sensitivity analysis
- Question: How large a departure from the SI 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

- **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:

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

Interpreting Sensitivity Analysis with R squares

- 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 parameters: R squares
  1. Proportion of previously unexplained variance explained by U_i

$$R^{2*}_M \equiv 1 - \frac{\text{var}(\epsilon'_{i2})}{\text{var}(\epsilon_{i2})} \quad \text{and} \quad R^{2*}_Y \equiv 1 - \frac{\text{var}(\epsilon'_{i3})}{\text{var}(\epsilon_{i3})}$$

  2. Proportion of original variance explained by U_i

$$\widetilde{R}^2_M \equiv \frac{\text{var}(\epsilon_{i2}) - \text{var}(\epsilon'_{i2})}{\text{var}(M_i)} \quad \text{and} \quad \widetilde{R}^2_Y \equiv \frac{\text{var}(\epsilon_{i3}) - \text{var}(\epsilon'_{i3})}{\text{var}(Y_i)}$$

- Then reparameterize ρ using (R_M^{2*}, R_Y^{2*}) (or $(\widetilde{R}_M^2, \widetilde{R}_Y^2)$):

$$\rho \quad = \quad \text{sgn}(\lambda_2\lambda_3) R_M^* R_Y^* \quad = \quad \frac{\text{sgn}(\lambda_2\lambda_3)\widetilde{R}_M\widetilde{R}_Y}{\sqrt{(1-R_M^2)(1-R_Y^2)}},$$

  where R_M^2 and R_Y^2 are from the original mediator and outcome models

- $\text{sgn}(\lambda_2\lambda_3)$ indicates the direction of the effects of U_i on Y_i and M_i

- Set (R_M^{2*}, R_Y^{2*}) (or $(\widetilde{R}_M^2, \widetilde{R}_Y^2)$) to different values and see how mediation effects change

Reanalysis: Estimates under Sequential Ignorability

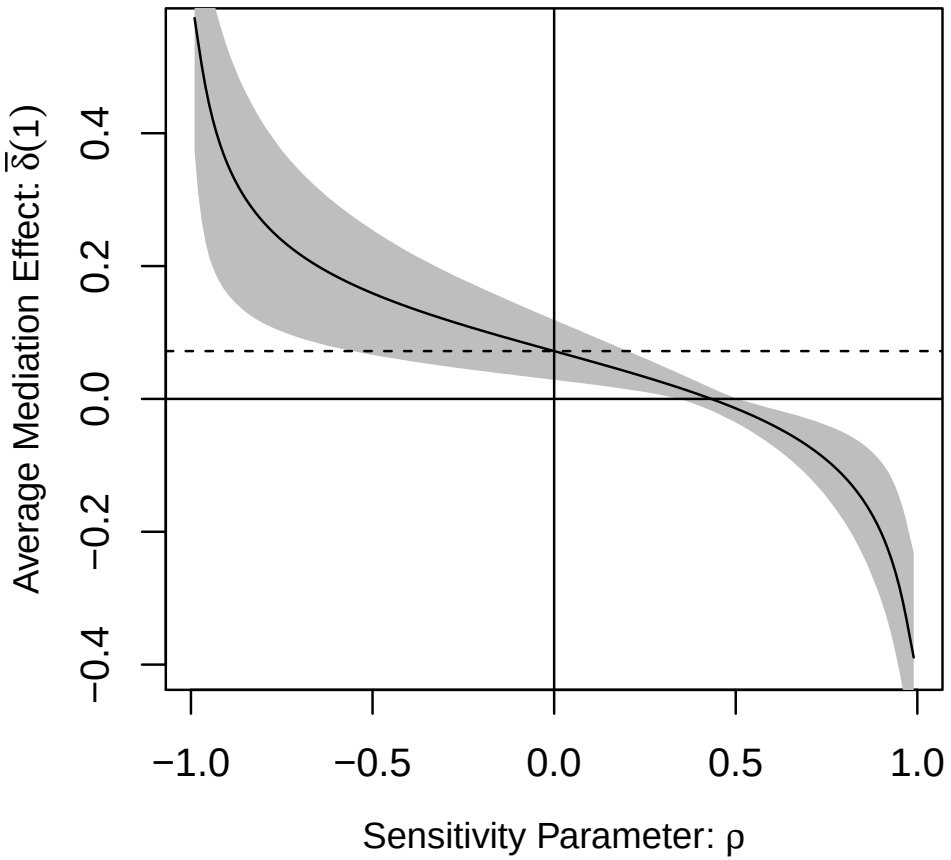
- Original method: Product of coefficients with the Sobel test

  — Valid only when both models are linear w/o T–M interaction (which they are not)

- Our method: Calculate ACME using our general algorithm

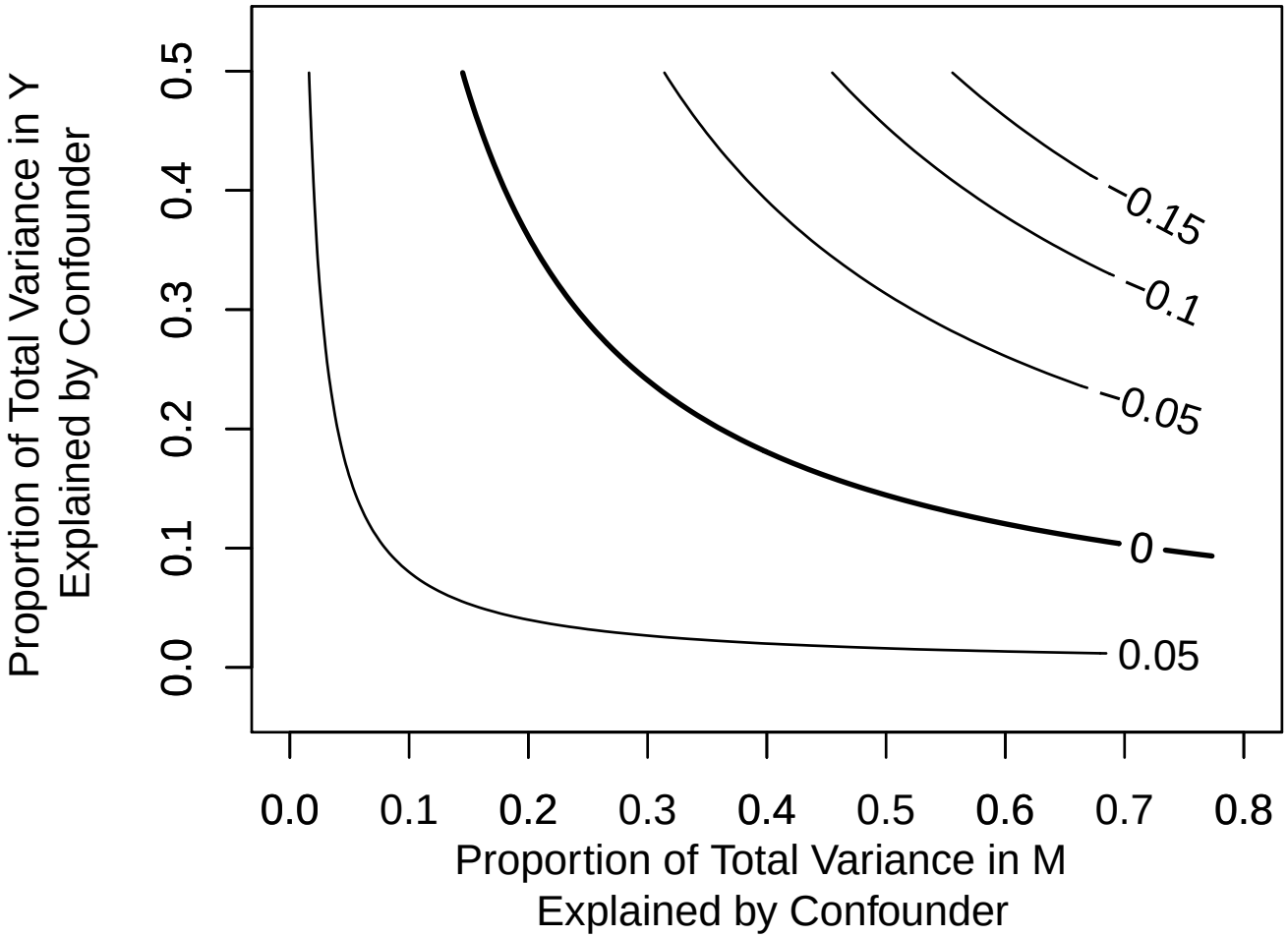
Outcome variables	Product of Coefficients	Average Causal Mediation Effect (δ)
Decrease Immigration	.347	.105
$\bar{\delta}(1)$	[0.146, 0.548]	[0.048, 0.170]
Support English Only Laws	.204	.074
$\bar{\delta}(1)$	[0.069, 0.339]	[0.027, 0.132]
Request Anti-Immigration Information	.277	.029
$\bar{\delta}(1)$	[0.084, 0.469]	[0.007, 0.063]
Send Anti-Immigration Message	.276	.086
$\bar{\delta}(1)$	[0.102, 0.450]	[0.035, 0.144]

Reanalysis: Sensitivity Analysis w.r.t. ρ



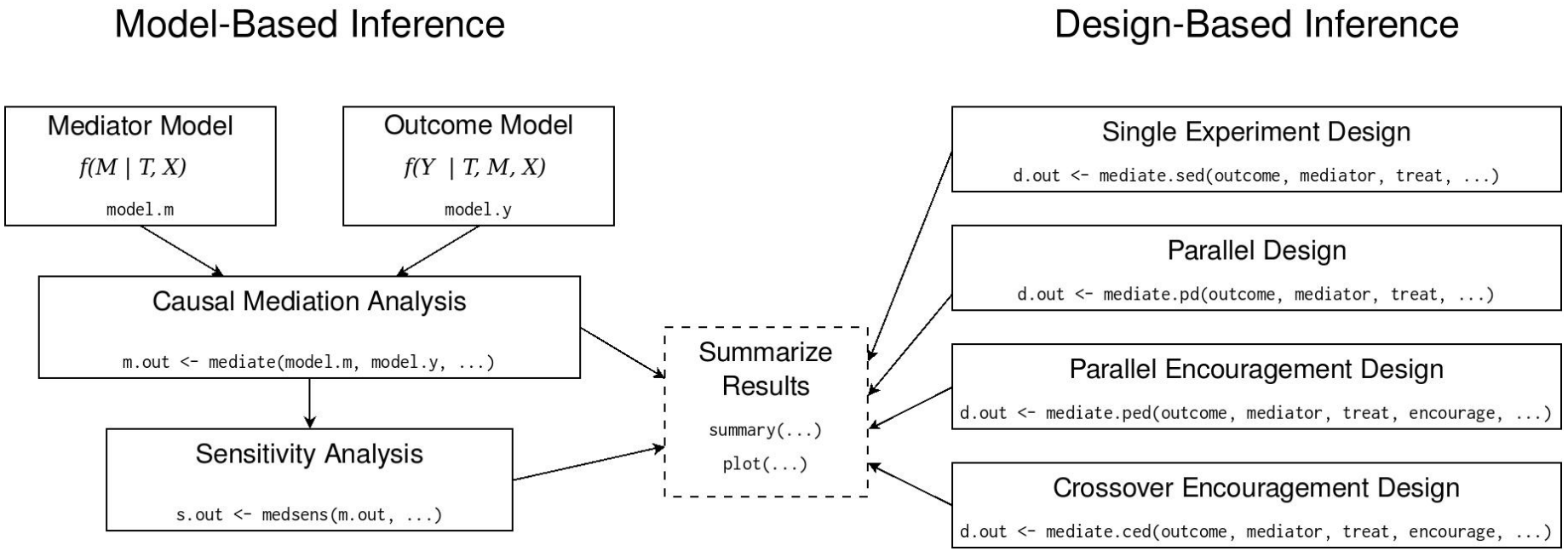
- ACME $>$ 0 as long as the error correlation is less than 0.39 (0.30 with 95% CI)

Reanalysis: Sensitivity Analysis w.r.t. $\tilde{R}^2_M$ 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

Overview of **R** Package **mediation**



- All of our proposed methods (and more!) can be implemented
- Stata version is also available but has limited capabilities

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")
```

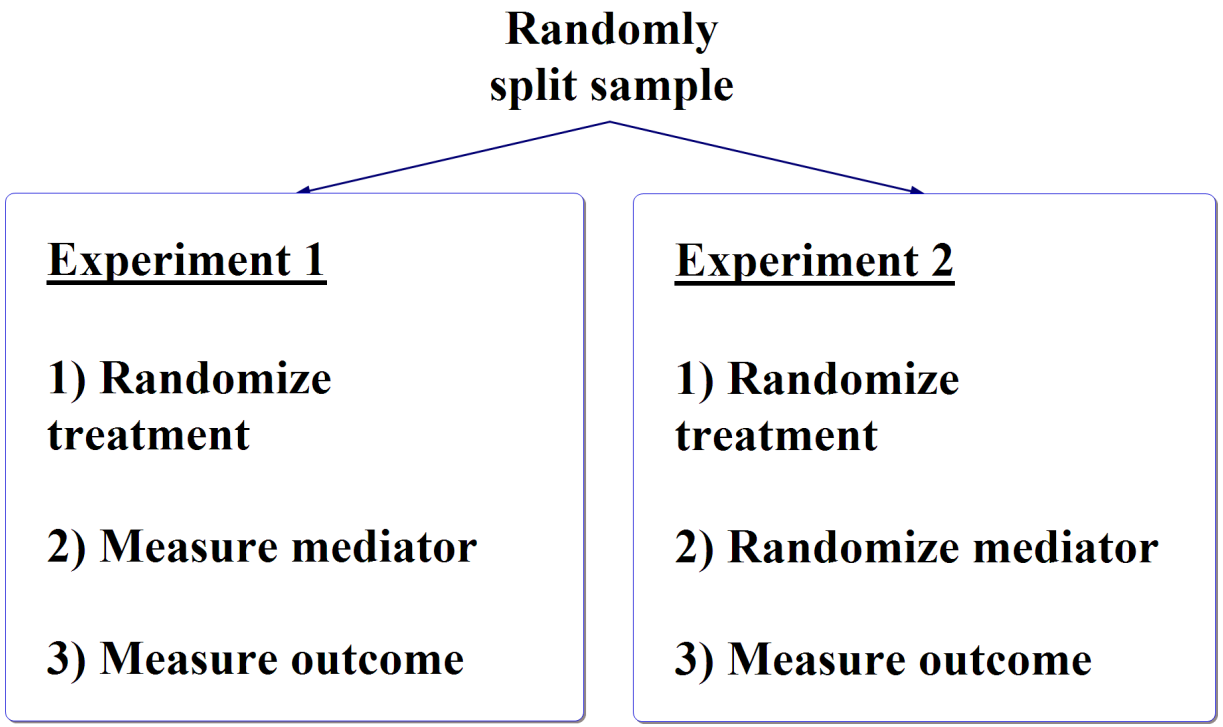
Beyond Sequential Ignorability

- 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

- All proposed designs preserve the ability to estimate the ACME under the SI assumption
- Trade-off: statistical power

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

- 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
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

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 doesn't 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
- Assumptions can be made plausible

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, propensity scores, etc.
- Search for quasi-randomized treatments: "natural" experiments

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

Cross-Over Design Examples from Political Science

EXAMPLE Incumbency advantage

- Estimation of incumbency advantages goes back to 1960s
- Why incumbency advantage? Scaring off quality challenger
- 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

The project website for papers and software:

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

Email for comments and suggestions:

kimai@Princeton.Edu