Introduction to Causal Mediation Analysis Using R

Teppei Yamamoto

Massachusetts Institute of Technology

ASA Webinar

March 9, 2017
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?
Webinar Overview

Present 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. Illustrate how to use **statistical software** mediation
5. Consider **research designs** that relax sequential ignorability (time permitting)
Causal Mediation Analysis

- Graphical representation

\[ \text{Mediator, } M \]

\[ \text{Treatment, } T \rightarrow \text{Outcome, } Y \]

- Goal is to decompose total effect into direct and indirect effects
- Alternative approach: decompose the treatment into different components
- Causal mediation analysis as \textit{quantitative process tracing}
Example: 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 NYT article 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.

Anxiety, $M$

Media Cue, $T$ → Immigration Attitudes, $Y$

- 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:

\[ Y_i = \alpha_1 + \beta_1 T_i + \xi_1^\top X_i + \epsilon_{1i} \]
\[ M_i = \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{2i} \]
\[ Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{3i} \]

where \( X_i \) is a set of pre-treatment or control variables.

1. Total effect (ATE) is \( \beta_1 \)
2. Direct effect is \( \beta_3 \)
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 \( \beta_1 \) 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 $i$
- $M_i(1)$: Level of anxiety individual $i$ would report if he read the story with Hispanic immigrant

- $Y_i(1, M_i(1))$: Immigration attitude individual $i$ would report if he read 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 (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 \( \implies \) Zero mediation effect

Example: Difference in immigration attitudes that is due to the change in anxiety induced by the treatment news story
Total Effect $= \text{Indirect Effect} + \text{Direct Effect}$

- **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** $= \text{mediation (indirect) effect} + \text{direct effect}$

- **cf. Controlled direct effects:**

\[
\xi_i(t, m, m') \equiv Y_i(t, m) - Y_i(t, m')
\]
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 $Y_i(t, M_i(1 - t))$

- 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** $\implies$ Need assumptions or better research designs
Counterfactuals in the Example

- Suppose that 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
- We can’t observe this because $M_i(0)$ is not realized when $T_i = 1$
Identification assumption: Sequential Ignorability (SI)

\[ \{ Y_i(t', m), M_i(t) \} \perp \perp T_i \mid X_i = x, \quad (1) \]
\[ Y_i(t', m) \perp \perp M_i(t) \mid T_i = t, X_i = x \quad (2) \]

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 \( \rightarrow \) 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
Back to Brader et al.:

- Treatment is randomized $\implies$ (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 $\implies$ measure and adjust for them
- Post-treatment confounders $\implies$ 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):**

  \[
  M_i = \alpha_2 + \beta_2 T_i + \xi_2^T X_i + \epsilon_{i2}, \\
  Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^T X_i + \epsilon_{i3}.
  \]

- 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
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. Output does not generally equal an interpretable effect size
A 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:

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

- Can’t multiply \(\beta_2\) by \(\gamma\)
- Difference of coefficients \(\beta_1 - \beta_3\) doesn’t work either

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

- 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 \(\Pr(Y_i(1, M_i(0)) = 1 \mid T_i = 1, M_i = \hat{M}_i(0), X_i)\) 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
Sensitivity parameter: \( \rho \equiv \text{Corr}(\epsilon_{i2}, \epsilon_{i3}) \)

Sequential ignorability implies \( \rho = 0 \)

Set \( \rho \) to different values and see how ACME changes

**Result:**

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

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?

\( \delta(t) \) = 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 $\rho$: 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 $\rho$ using $(R_{M}^{2*}, R_{Y}^{2*})$ (or $(\tilde{R}_{M}^{2}, \tilde{R}_{Y}^{2})$):

$$\rho = \text{sgn}(\lambda_2 \lambda_3) R_{M}^{*} R_{Y}^{*} = \frac{\text{sgn}(\lambda_2 \lambda_3) \tilde{R}_{M} \tilde{R}_{Y}}{\sqrt{1 - R_{M}^{2}}(1 - R_{Y}^{2})},$$

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 $(\tilde{R}_{M}^{2}, \tilde{R}_{Y}^{2})$) to different values and see how mediation effects change
Reanalysis of Brader et al.: Estimates under SI

- 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

<table>
<thead>
<tr>
<th>Outcome variables</th>
<th>Product of Coefficients</th>
<th>Average Causal Mediation Effect ($\delta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decrease Immigration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta(1)$</td>
<td>0.347</td>
<td>0.105</td>
</tr>
<tr>
<td>[0.146, 0.548]</td>
<td>[0.048, 0.170]</td>
<td></td>
</tr>
<tr>
<td>Support English Only Laws</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta(1)$</td>
<td>0.204</td>
<td>0.074</td>
</tr>
<tr>
<td>[0.069, 0.339]</td>
<td>[0.027, 0.132]</td>
<td></td>
</tr>
<tr>
<td>Request Anti-Immigration Information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta(1)$</td>
<td>0.277</td>
<td>0.029</td>
</tr>
<tr>
<td>[0.084, 0.469]</td>
<td>[0.007, 0.063]</td>
<td></td>
</tr>
<tr>
<td>Send Anti-Immigration Message</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta(1)$</td>
<td>0.276</td>
<td>0.086</td>
</tr>
<tr>
<td>[0.102, 0.450]</td>
<td>[0.035, 0.144]</td>
<td></td>
</tr>
</tbody>
</table>
ACME > 0 as long as the error correlation is less than 0.39 (0.30 with 95% CI)
Reanalysis: Sensitivity Analysis w.r.t. $\tilde{R}_M^2$ and $\tilde{R}_Y^2$

An unobserved confounder can account for up to 26.5% of the variation in both $Y_i$ and $M_i$ before ACME becomes zero.
What does it mean when the mediation effect has a different sign from the total effect?

I don’t understand the difference between $\delta_i(0)$ and $\delta_i(1)$.

Do I always have to measure the mediator before the outcome?

My treatment is continuous. How do I choose values of $t$ and $t'$?
Q. I got an ACME that was the opposite of the total effect, what does that mean?

A. Recall the identity: Total Effect = ACME + Direct Effect. Therefore, ACME and direct effects must have opposite signs and the direct effect is larger in magnitude.

**Example**  
\[ T = \text{job training}, \ Y = \text{earnings}, \ M = \text{skills} \]

Suppose: Total effect \(< 0\) and ACME \(> 0\)

It must be the case: Direct effect \(<< 0\)

That is, there must be some other mechanism (e.g. time spent without searching for jobs) which is more important (quantitatively) than improved skills and makes the net impact of job training on earnings negative.
Q. I don’t understand the difference between $\delta_i(0)$ and $\delta_i(1)$. When is one more important than the other?

One can relax the so-called no interaction rule with the following model for the outcome:

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \kappa T_i M_i + \xi_3^T X_i + \epsilon_{i3}.$$ for $t = 0, 1$. The average causal mediation effects are given by,

$$\bar{\delta}(t) = \beta_2 (\gamma + \kappa t),$$
Q. I don’t understand the difference between $\delta_i(0)$ and $\delta_i(1)$. When is one more important than the other?

A. The difference is which condition is considered *actual* and which is *counterfactual*.

$\delta_i(0)$: The effect that the treatment would have had if its only action were to cause the mediator. (Actual world = control)

$\delta_i(1)$: The effect of treatment that would be prevented if the exposure did not cause the mediator. (Actual world = treated)

Oftentimes the control condition represents the “natural” state of the world or a “status quo.” In this case $\delta_i(0)$ may be the more relevant quantity.

Epidemiologists sometimes call $\delta_i(0)$ the *pure indirect effect* for this reason.
Q. Do I always have to measure the mediator before the outcome?

A. Yes, unless you have a really good reason to believe that measuring the outcome has no effect (or only has a negligibly small effect) on the measurement of the mediator.

Even if the mediator cannot be affected by the outcome conceptually, the measurement error in the mediator (which is unavoidable in most cases) can be affected by the outcome, contaminating the estimates. This is a measurement error problem much broader than mediation analysis (see Imai and Yamamoto 2010 AJPS).
Q. My treatment is continuous. How do I choose values of $t$ and $t'$?

A. There are several sensible ways to approach this problem:

1. If there are two values that are substantively interesting (e.g. correspond to the two most typical values in the real world), use them.

2. If the empirical distribution of the treatment is bimodal, use two values that represent the two modes.

3. If there is one value that can be regarded as a “baseline” (e.g. no treatment, natural condition), use that value as $t'$, compute multiple ACMEs by setting $t$ to many different values, and plot the estimates against $t$.

4. If there is a natural “cutpoint” in the treatment values, dichotomize the treatment variable before the estimation and treat it as a binary variable (i.e. high vs. low).
Model-Based Inference

Mediator Model
\[ f(M \mid T, X) \]
\[ \text{model.m} \]

Outcome Model
\[ f(Y \mid T, M, X) \]
\[ \text{model.y} \]

Causal Mediation Analysis
\[ \text{m.out} \leftarrow \text{mediate(model.m, model.y, ...)} \]

Sensitivity Analysis
\[ \text{s.out} \leftarrow \text{medsens(m.out, ...)} \]

Design-Based Inference

Single Experiment Design
\[ \text{d.out} \leftarrow \text{mediate.sed(outcome, mediator, treat, ...)} \]

Parallel Design
\[ \text{d.out} \leftarrow \text{mediate.pd(outcome, mediator, treat, ...)} \]

Parallel Encouragement Design
\[ \text{d.out} \leftarrow \text{mediate.ped(outcome, mediator, treat, encourage, ...)} \]

Crossover Encouragement Design
\[ \text{d.out} \leftarrow \text{mediate.ced(outcome, mediator, treat, encourage, ...)} \]
Implementation Examples

1. Fit models for the mediator and outcome variable and store these models
   
   ```r
   > 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
   
   ```r
   > 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
   
   ```r
   > s.out <- medsens(m.out)
   > summary(s.out)
   > plot(s.out, "rho")
   > plot(s.out, "R2")
   ```
## Data Types Available via mediation

<table>
<thead>
<tr>
<th>Mediator Model Types</th>
<th>Outcome Model Types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
</tr>
<tr>
<td>Linear (lm/lmer)</td>
<td>✓</td>
</tr>
<tr>
<td>GLM (glm/bayesglm/glmer)</td>
<td>✓</td>
</tr>
<tr>
<td>Ordered (polr/bayespolr)</td>
<td>✓</td>
</tr>
<tr>
<td>Censored (tobit via vglm)</td>
<td>-</td>
</tr>
<tr>
<td>Quantile (rq)</td>
<td>✓*</td>
</tr>
<tr>
<td>GAM (gam)</td>
<td>✓*</td>
</tr>
<tr>
<td>Survival (survreg)</td>
<td>✓</td>
</tr>
</tbody>
</table>

Types of Models That Can be Handled by `mediate`. Stars (*) indicate the model combinations that can only be estimated using the nonparametric bootstrap (i.e. with `boot = TRUE`).
Additional Features

- Treatment/mediator interactions, with formal statistical tests
- Treatment/mediator/pre-treatment interactions and reporting of quantities by pre-treatment values
- Factoral, continuous treatment variables
- Cluster standard errors/adjustable CI reporting/p-values
- Support for multiple imputation
- Multiple mediators
- Multilevel mediation

A tutorial with examples: Tingley et al. (2014) (click and download).
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

- These experimental designs can then be extended to natural experiments in observational studies
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:

<table>
<thead>
<tr>
<th>Prop.</th>
<th>$M_i(1)$</th>
<th>$M_i(0)$</th>
<th>$Y_i(t, 1)$</th>
<th>$Y_i(t, 0)$</th>
<th>$\delta_i(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- $\mathbb{E}(M_i(1) - M_i(0)) = \mathbb{E}(Y_i(t, 1) - Y_i(t, 0)) = 0.2$, 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”
Encouragement Design

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

\[ Z \rightarrow M \]
\[ T \rightarrow Y \]
\[ M \rightarrow Y \]

- 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)
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
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
Multiple Mediators

- Quantity of interest = The average indirect effect with respect to $M$
- $W$ represents the alternative observed mediators
- Left: Assumes independence between the two mechanisms
- Right: Allows $M$ to be affected by the other mediators $W$
- Applied work often assumes the independence of mechanisms
- Under this independence assumption, one can apply the same analysis as in the single mediator case
- For causally dependent mediators, we must deal with the heterogeneity in the $T \times M$ interaction as done under the parallel design $\rightarrow$ sensitivity analysis
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.
Thank you!

Collaborators on the project:

- Kosuke Imai (Princeton)
- Luke Keele (Georgetown)
- Dustin Tingley (Harvard)
- Kentaro Hirose (Waseda)

Send your questions and comments to:

teppei@mit.edu
References (click on the article titles)

- **General:**
  - Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review*

- **Theory:**

- **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*