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

Kosuke Imai Luke Keele
Dustin Tingley Teppei Yamamoto

APSA Short Course

August 28, 2013

Project References (click the article titles)

- **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.
  - Causal Mediation Analysis Using R.

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?

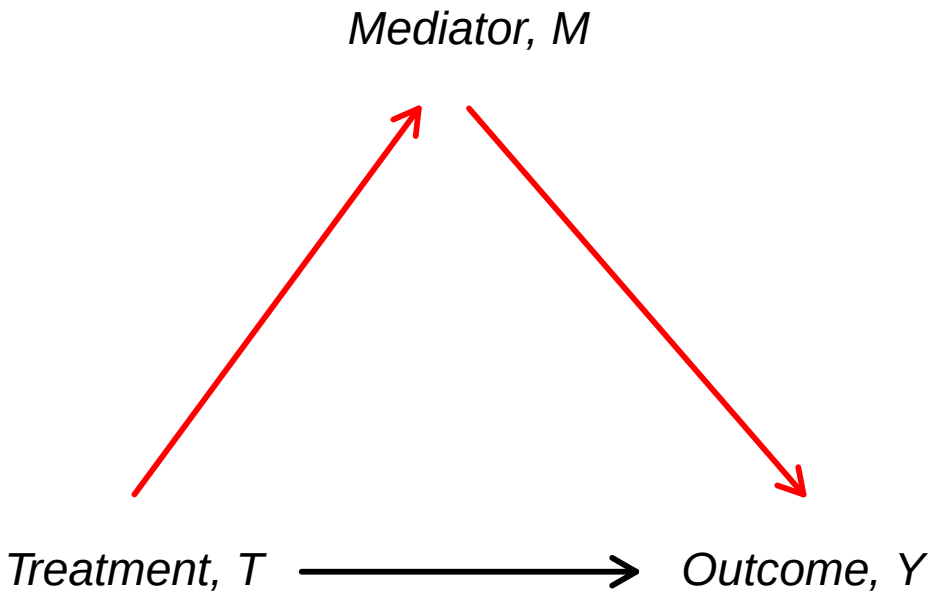
Goals of the Course

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 R package `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

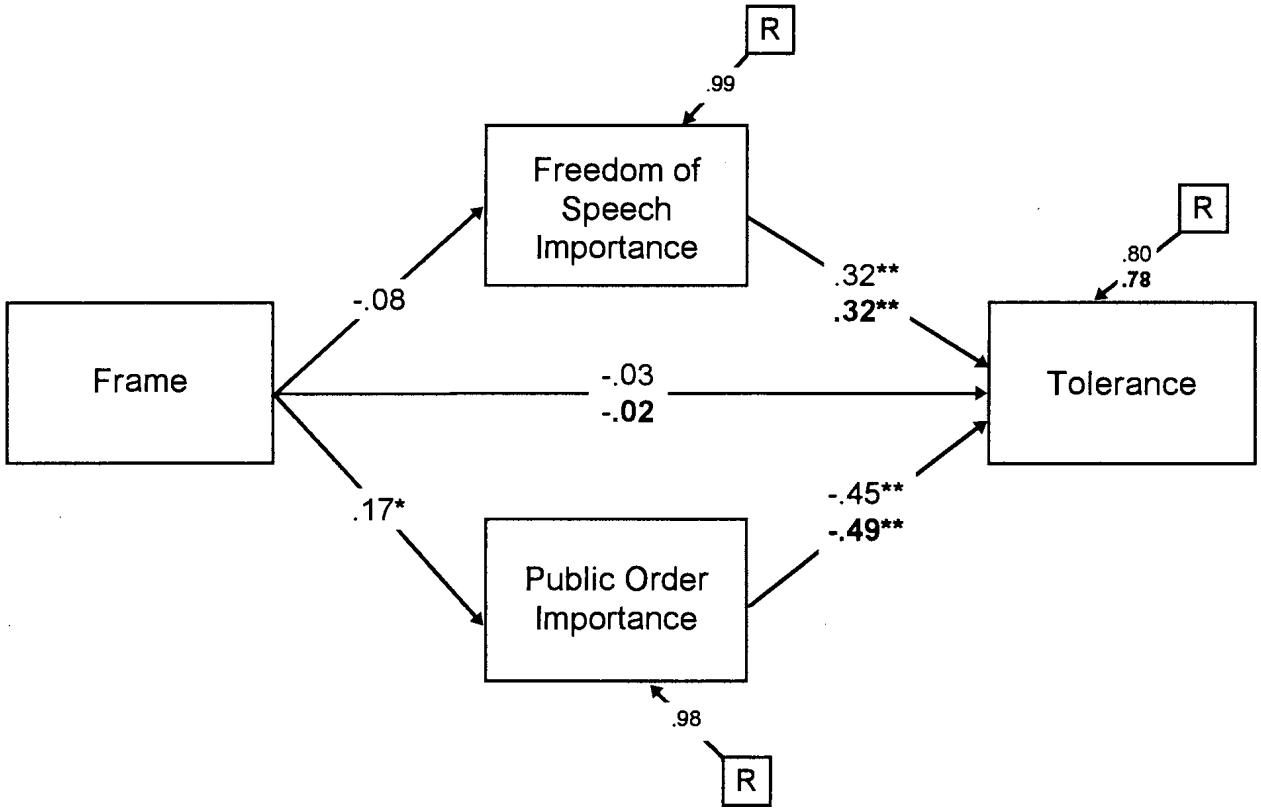
- Graphical representation



- Goal is to decompose total effect into direct and indirect effects.

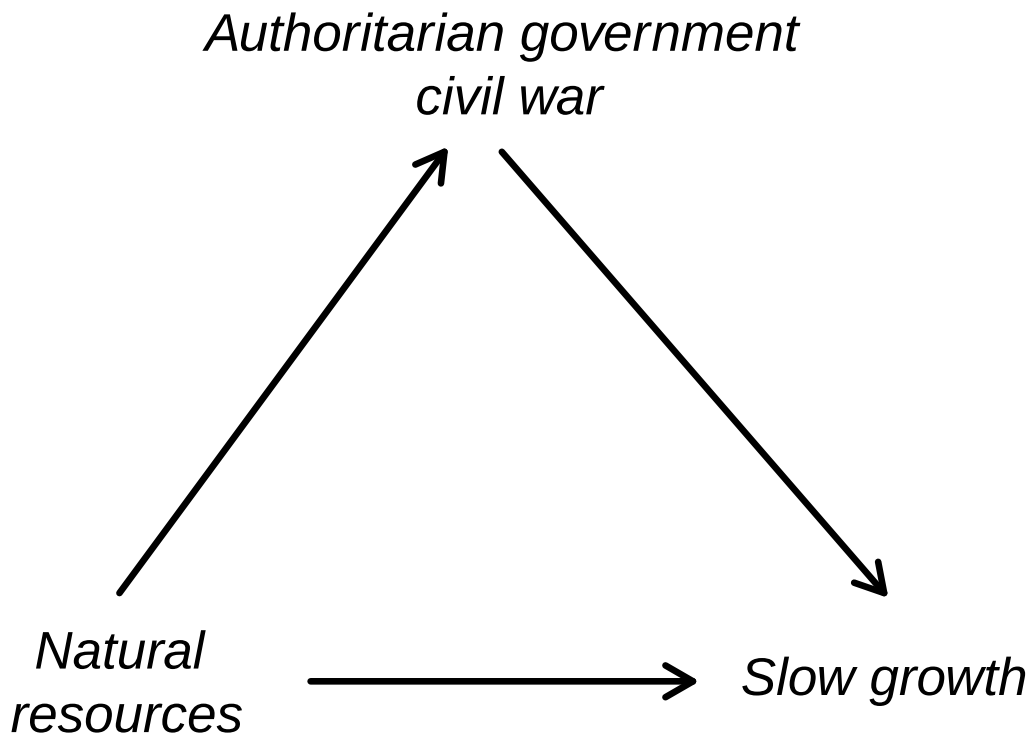
Causal Mediation Analysis in **American Politics**

- The political psychology literature on media framing.
- Nelson *et al*. (*APSR*, 1998)



Causal Mediation Analysis in **Comparative Politics**

- Resource curse thesis



- Causes of civil war: Fearon and Laitin (*APSR*, 2003)

Causal Mediation Analysis in **International Relations**

- The literature on international regimes and institutions
- Krasner (*International Organization*, 1982)

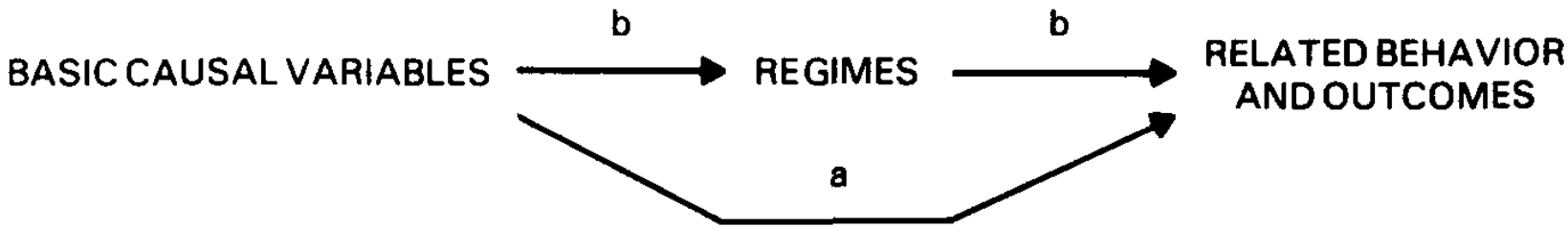


Figure 2

- Power and interests are mediated by regimes

Standard Estimation Methods

Standard Equations for Mediator and Outcome:

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

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

But what must we assume for the decomposition to represent causal effects?

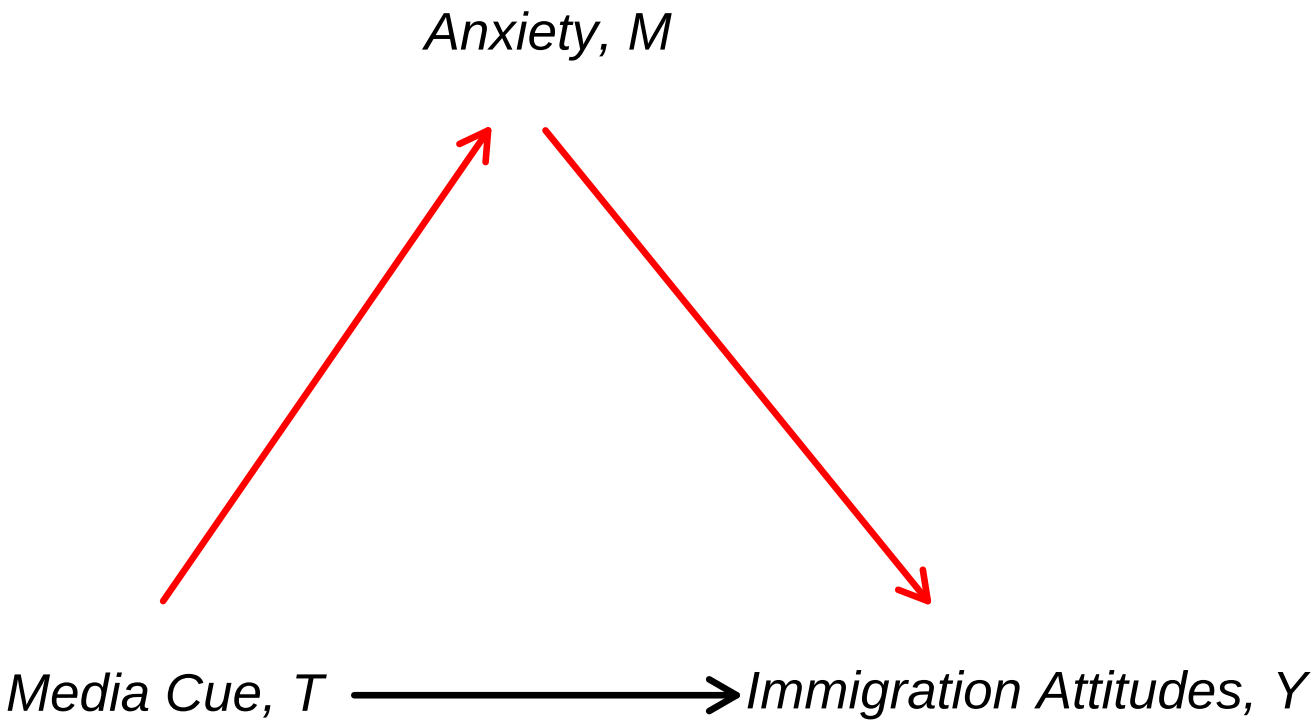
Media Cues and Immigration Attitudes

Brader et al. experiment:

- Subjects read a mock news story about immigration.
- Treatment: immigrant in story is a Hispanic, and the news story emphasized the economic costs of immigration.
- They measured a range of different attitudinal and behavioral outcome variables:
  - Opinions about increasing or decrease immigration,
  - Contact legislator about the issue,
  - Send anti-immigration message to legislator...

They want to test whether the treatment increases anxiety, leading to greater opposition to immigration.

Causal Mediation Analysis in Brader et al.



What is the effect of the news story that works through making people anxious?

Let's translate this theory into counterfactual quantities.

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

Example with This Notation

$M_i(1)$ is the observed level of anxiety reported by individual i, who was assigned to the treatment condition (read negative story with Hispanic immigrant).

$Y_i = Y_i(1, M_i(1))$ is the observed immigration attitude reported by individual i, who was assigned to the treatment condition (read negative story with Hispanic immigrant), and had the observed anxiety level $M_i(1)$.

$M_i(0)$ and $Y_i = Y_i(0, M_i(0))$ are the converse.

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
- In the Brader example: 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:

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

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: $\xi(1, m, m') - \xi(0, m, m') \neq 0$
- Doesn't imply the existence of a mechanism

What Does the Data Tell Us?

- Recall the Brader et al. experimental design: randomize T_i, measure M_i and Y_i.
- $Y_i = Y_i(t, M_i(\mathbf{t}))$ is observed but not $Y_i = Y_i(t, M_i(\mathbf{1\text{-}t}))$
- But we want to estimate

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$

  for $t = 0, 1$, which is not directly in the data.

- What is this counterfactual potential outcome?

The Counterfactual

- Think of a subject that viewed the treatment news story ($t_i = 1$).
- For this person, $Y_i(1, M_i(1))$ is the observed immigration opinion if he or she views the immigration news story.
- $Y_i(1, M_i(0))$ is his or her immigration opinion in the counterfactual world where subject *i* still viewed the immigration story but his or her anxiety level is at the same level as if they viewed the control news story.
- We face an "identification problem" since we don't observe $Y_i(1, M_i(0))$

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

Under sequential ignorability, both 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 $\bar{\zeta}(t)$

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

Sequential Ignorability in the Brader Example

- Brader et al looked at two different mediators or mechanisms.
- One is anxiety.
- Second is the participants' belief about the likely negative impact of immigration what they called perceived harm.
- Easy to think of confounders for this mechanism.
- One could be state. Those who live in AZ are more likely to have higher levels of perceived harm and more likely to be opposed to immigration.
- One must measure and control for all possible confounders that could affect both mediator and outcome.

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
- Problem: Often only the test of significance is reported

Traditional Estimation Methods: “Baron-Kenny Steps”

Argument: you must show all three steps in your paper in order to claim a significant relationship

- Regress Y on T and show there is a significant relationship
- Regress M on T and show there is a significant relationship
- Regress Y on M and T, and show there is a significant relationship between M and Y

Problems

- First step can lead to false negatives if indirect and direct effects in different directions
- Does quantify the thing you’re trying to quantify. Looking at stars is silly.
- Proponents only used it for linear regression models, and well documented its of lower power compared to LSEM.

Don’t do this! Instead...

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: Continuous Mediator and Binary Outcome

Estimate the two following models:

$$
\begin{aligned}
M_i &= \alpha_2 + \beta_2 T_i + X_i + \epsilon_{2i}, \\
\Pr(Y_i = 1) &= \Phi\left(\alpha_3 + \beta_3 T_i + \gamma M_i + X_i + \epsilon_{3i}\right)
\end{aligned}
$$

- Predict M_i for $T_i = 1$ and $T_i = 0$. This gives you $\hat{M}_i(1)$ and $\hat{M}_i(0)$.
- Predict Y_i with $T_i = 1$ and $\hat{M}_i(0)$ and vice versa.
- Take average of these two predictions.

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

- **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

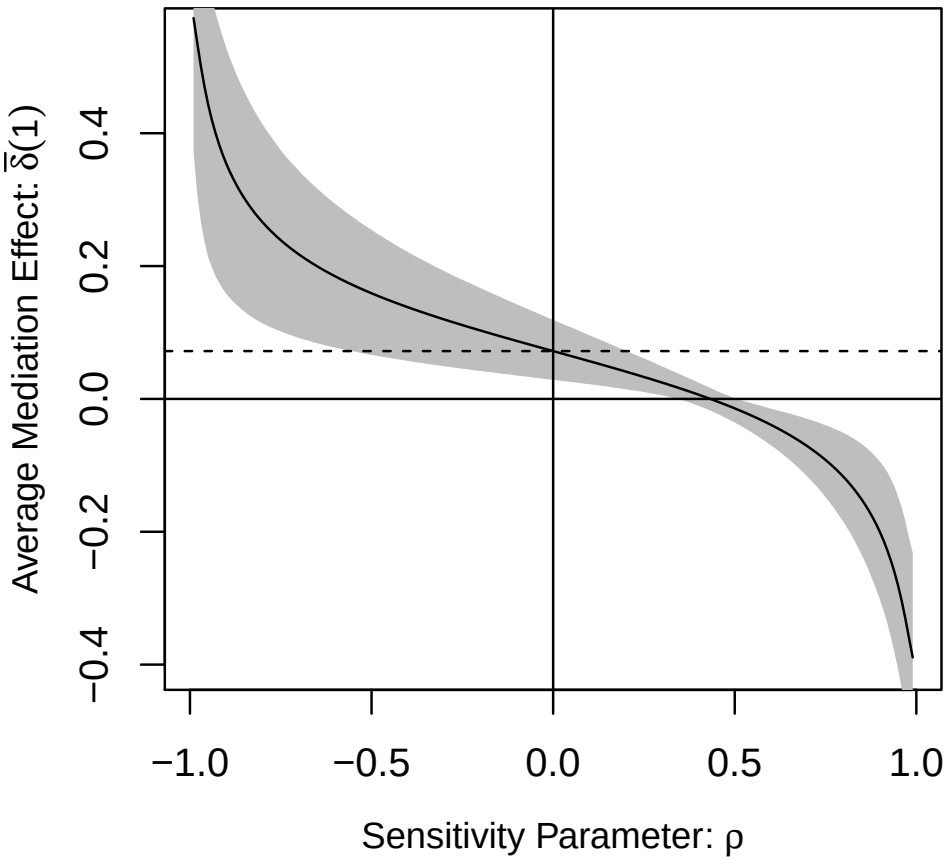
Outcome variables	Product of Coefficients	Average Causal Mediation Effect (δ)
Decrease Immigration	.347	.105
$\bar{\delta}(1)$	[0.146, 0.548]	[0.048, 0.170]
Support English Only Laws	.204	.074
$\bar{\delta}(1)$	[0.069, 0.339]	[0.027, 0.132]
Request Anti-Immigration Information	.277	.029
$\bar{\delta}(1)$	[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]

Special Focus: Binary Outcomes

How do I interpret the indirect effect when outcome is binary?

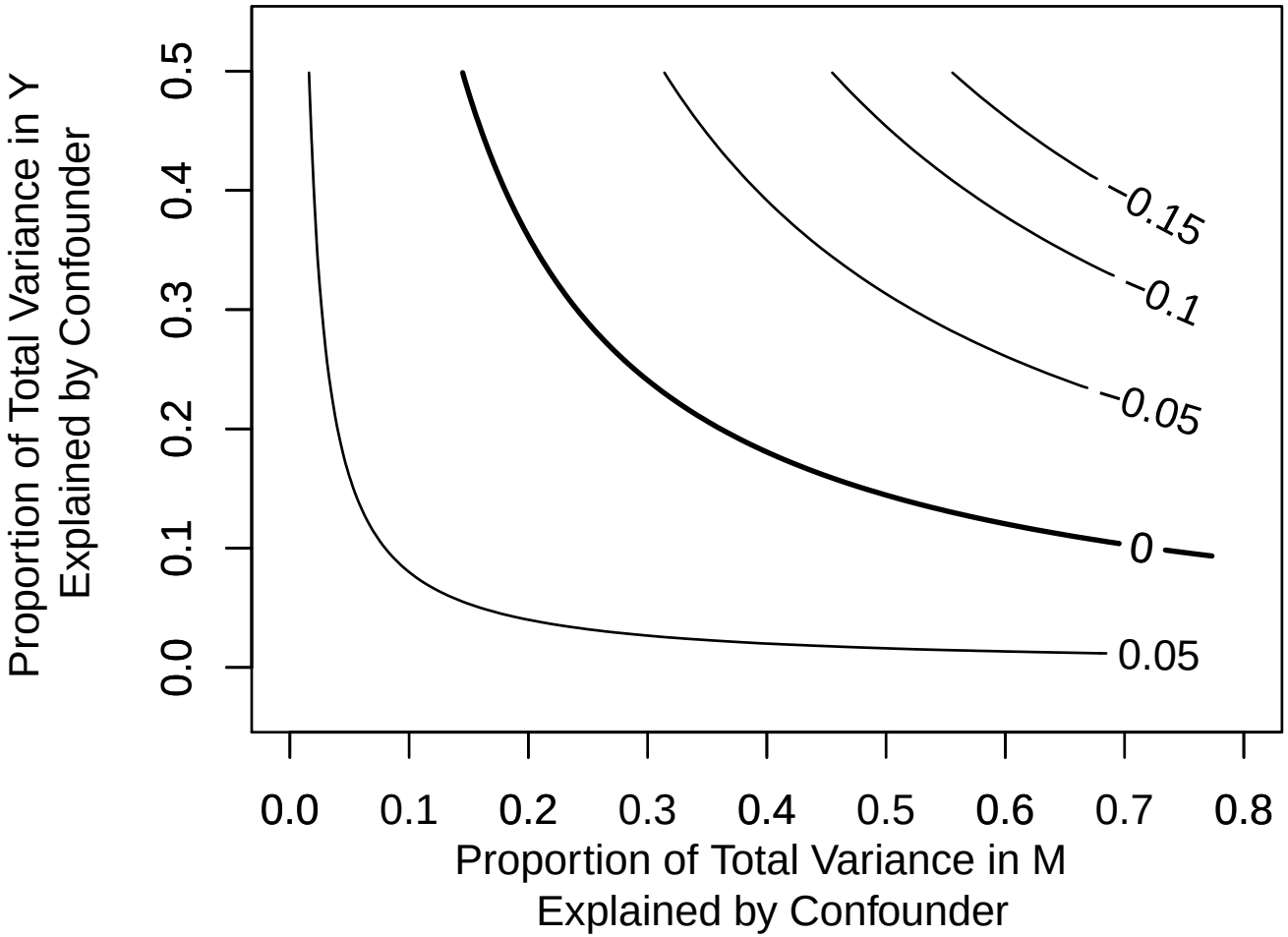
	Product of Coefficients	Average Causal Mediation Effect (δ)
Send Anti-Immigration Message $\bar{\delta}(1)$	.276 [0.102, 0.450]	.086 [0.035, 0.144]

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

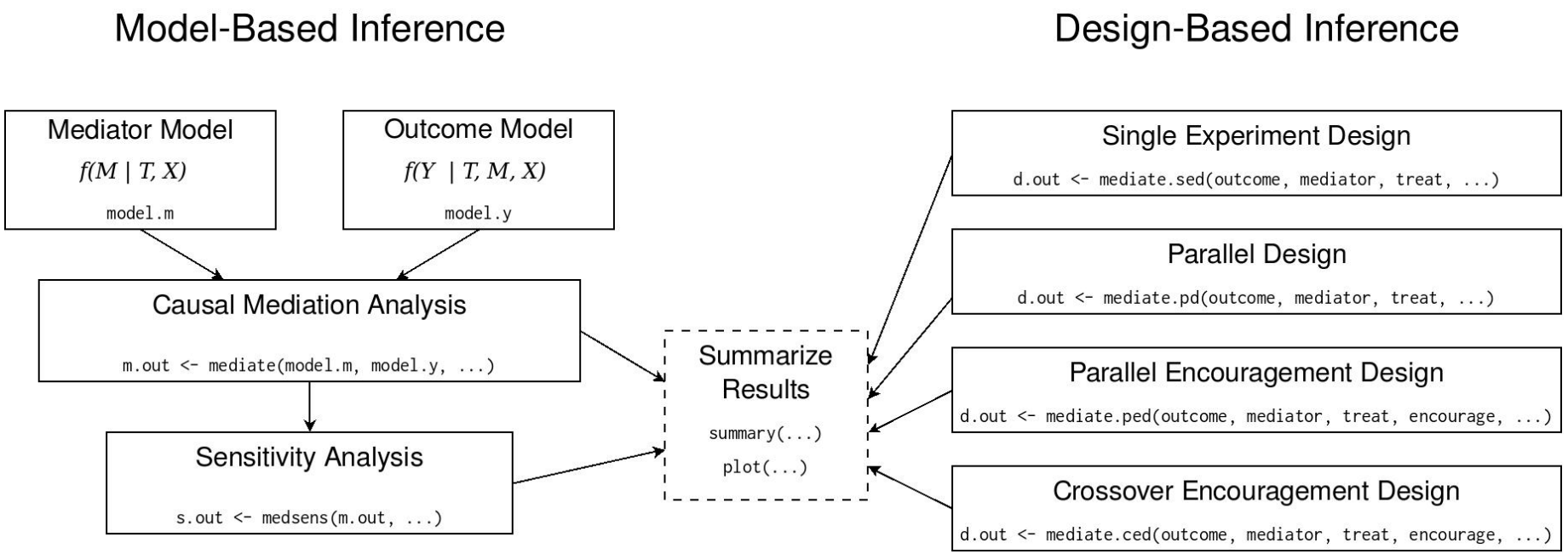


Figure: Structure of the **R mediation** package as of version 4.0.

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

```
> m <- lm(Mediator ~ Treat + X,data=Data)
> y <- lm(Outcome ~ Treat + Mediator + X,data=Data)
```

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")
```

4. Experimental designs and analysis now also available

Data Types Available via **mediation**

	Outcome Model Types						
Mediator Model Types	Linear	GLM	Ordered	Censored	Quantile	GAM	Survival
Linear (`lm`/`lmer`)	✓	✓	✓*	✓	✓	✓*	✓
GLM (`glm`/`bayesglm`/`glmer`)	✓	✓	✓*	✓	✓	✓*	✓
Ordered (`polr`/`bayespolr`)	✓	✓	✓*	✓	✓	✓*	✓
Censored (tobit via `vglm`)	-	-	-	-	-	-	-
Quantile (`rq`)	✓*	✓*	✓*	✓*	✓*	✓*	✓
GAM (`gam`)	✓*	✓*	✓*	✓*	✓*	✓*	✓*
Survival (`survreg`)	✓	✓	✓*	✓	✓	✓*	✓

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 (**NEW!**)

Please read our vignette file **here**.

Data Types Available for Sensitivity Analysis

	Outcome		
Mediator	Continuous	Ordered	Binary
Continuous	Yes	No	Yes
Ordered	No	No	No
Binary	Yes	No	No

Causal Mediation Analysis in Stata

Based on the same algorithm

Hicks, R, Tingley D. 2011. Causal Mediation Analysis. Stata Journal. 11(4):609-615.

```
ssc install mediation
```

More limited coverage of models (just bc. of time though!)

Syntax: medeff

```
medeff (equation 1) (equation 2) [if] [in] [[weight]] ,
[sims(integer) seed(integer) vce(vcetype) Level(#)
interact(varname)] mediate(varname) treat(varname)
```

Where “equation 1” or “equation 2” are of the form (For equation 1, the mediator equation):

```
probit M T x
```

or

```
regress M T x
```

FAQs

- 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 $T =$ oil, $Y =$ growth, $M =$ authoritarianism

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. civil war) which is more important (quantitatively) than authoritarianism and makes the net impact of oil on growth 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:

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \kappa T_i M_i + \xi_3^\top X_i + \epsilon_{i3}.$$

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:

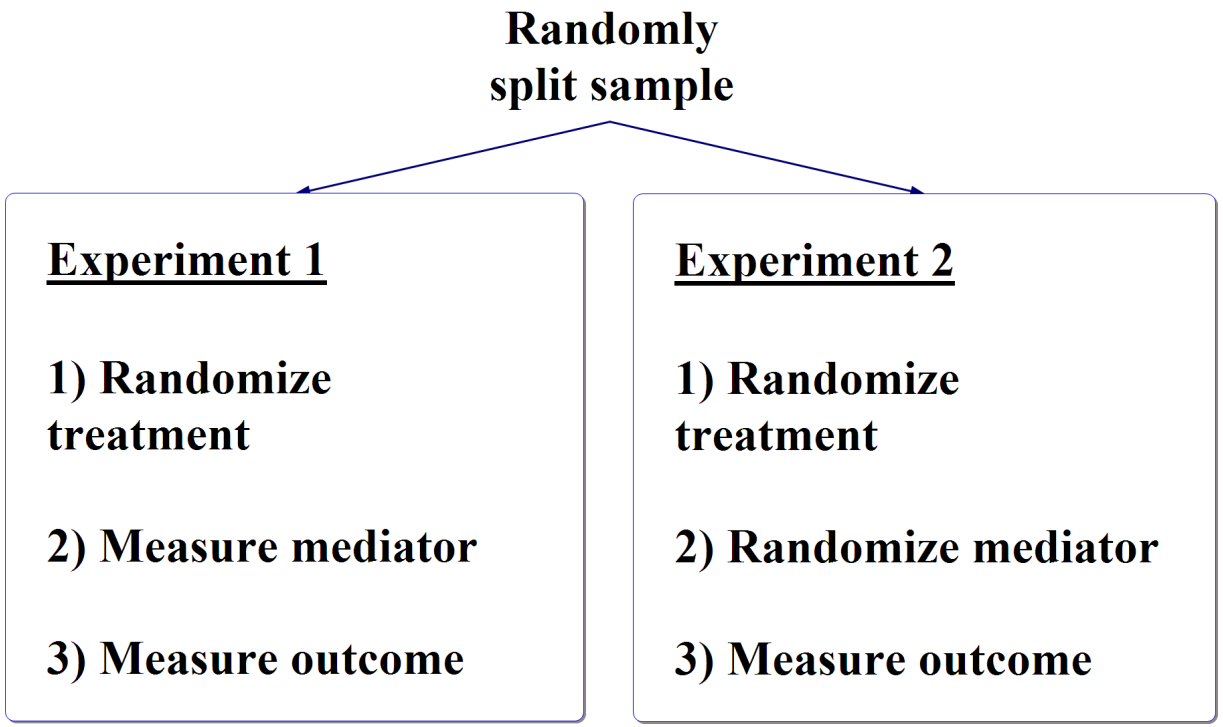
1. If there are two values that are substantively interesting (e.g. correspond to the two most typical values in the real world), use them.
2. If the empirical distribution of the treatment is bimodal, use two values that represent the two modes.
3. If there is one value that can be regarded as a "baseline" (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).

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



- 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

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
- 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

- Assumptions are plausible

Crossover Encouragement Design

- Crossover encouragement design:
  1. Round 1: Conduct a standard experiment
  2. Round 2: Same as crossover, except encourage subjects to take the mediator values

EXAMPLE Hainmueller & Hiscox (2010, APSR)

- Treatment: Framing immigrants as low or high skilled
- Outcome: Preferences over immigration policy
- Possible mechanism: Low income subjects may expect higher competition from low skill immigrants

- Manipulate expectation using a news story
- Round 1: Original experiment but measure expectation
- Round 2: Flip treatment, but encourage expectation in the same direction as Round 1

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

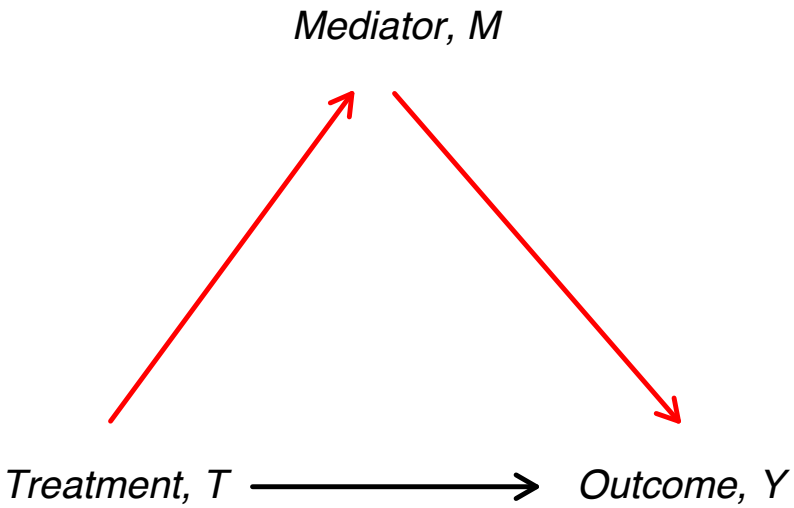
Example from Political Science

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

Extension to Multiple Mediators

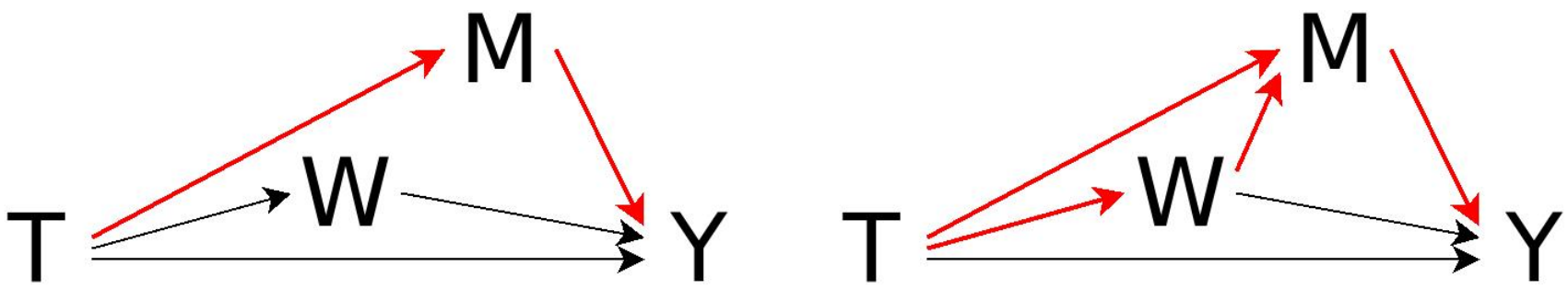
- Existing work typically focuses on a single mechanism:



  - How much of the treatment effect goes through M?
  - Potential outcomes framework
  - Total effect $=$ indirect effect $+$ direct effect

- However, multiple mediators are common in applied settings

Causally Independent vs. Dependent Mechanisms



- 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*
- Note that *W* can also be seen as post-treatment confounders between *M* and *Y*

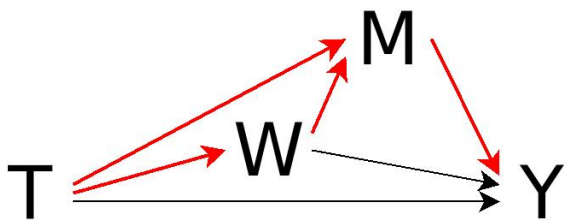
- Applied work often assumes the independence of mechanisms

Causally Related Multiple Mechanisms

- Binary treatment: $T_i \in \{0, 1\}$
- We allow W to influence both M and Y:

Potential mediators: $W_i(t)$ and $M_i(t, w)$

Potential outcomes: $Y_i(t, m, w)$



- Causal mediation effect (natural indirect effect):

$$\delta_i(t) \equiv Y_i(t, M_i(1, W_i(1)), W_i(t)) - Y_i(t, M_i(0, W_i(0)), W_i(t))$$

- Causal effect of the change in M_i induced by T_i
- Natural direct effect:

$$\zeta_i(t) \equiv Y_i(1, M_i(t, W_i(t)), W_i(1)) - Y_i(0, M_i(t, W_i(t)), W_i(0))$$

- Causal effect of T_i on Y_i holding M_i at its natural value when $T_i = t$
- These sum up to the total effect (as in the single mediator case)

Identification of Causally Related Mechainsms

- The FRCISTG assumption (Robins 1986):

$$\{Y_i(t,m,w), M_i(t,w), W_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x$$
$$\{Y_i(t,m,w), M_i(t,w)\} \perp\!\!\!\perp W_i \mid T_i = t,\ X_i = x$$
$$\{Y_i(t,m,w)\} \perp\!\!\!\perp M_i \mid W_i(t) = w,\ T_i = t,\ X_i = x$$

- A weak version of the sequential ignorability assumption
- Observed posttreatment confounding (W) is allowed (cf. Imai et al. 2010)
- Empirically verifiable, at least in theory

- Robins (2003): Under FRCISTG, the no interaction assumption nonparametrically identifies $\bar{\delta}(t)$:

$$Y_i(1,m,W_i(1)) - Y_i(0,m,W_i(0)) = Y_i(1,m',W_i(1)) - Y_i(0,m',W_i(0))$$

Allowing Interactions with Varying Coefficient LSEM

- Problem: The no interaction assumption is too strong in most cases
  (e.g. Is the effect of issue importance invariant across frames?)
- Solution: Assume a flexible model

$$
\begin{aligned}
M_i(t, w) &= \alpha_2 + \beta_{2i} t + \xi_{2i}^\top w + \mu_{2i}^\top tw + \lambda_{2i}^\top x + \epsilon_{2i}, \\
Y_i(t, m, w) &= \alpha_3 + \beta_{3i} t + \gamma_i m + \kappa_i tm + \xi_{3i}^\top w + \mu_{3i}^\top tw + \lambda_{3i}^\top x + \epsilon_{3i},
\end{aligned}
$$

  where $\mathbb{E}(\epsilon_{2i}) = \mathbb{E}(\epsilon_{3i}) = 0$

- Allows for dependence of M on W
- Coefficients can vary arbitrarily across units (= heterogeneous effects)

Sensitivity Analysis w.r.t. Interaction Heterogeneity

- The model can be rewritten as:

$$\begin{aligned} M_i(t, w) &= \alpha_2 + \beta_2 t + \xi_2^\top w + \mu_2^\top tw + \lambda_2^\top x + \eta_{2i}(t, w), \\ Y_i(t, m, w) &= \alpha_3 + \beta_3 t + \gamma m + \kappa tm + \xi_3^\top w + \mu_3^\top tw + \lambda_3^\top x + \eta_{3i}(t, m, w), \end{aligned}$$

  where $\beta_2 = \mathbb{E}(\beta_{2i})$, etc.

- FRCISTG implies

$$\mathbb{E}(\eta_{2i}(T_i, W_i) \mid X_i, T_i, W_i) = \mathbb{E}(\eta_{3i}(T_i, M_i, W_i) \mid X_i, T_i, W_i, M_i) = 0$$

  The mean coefficients β_2, etc. can thus be estimated without bias

- We can show that $\bar{\delta}(t)$ and $\bar{\zeta}(t)$ can be written as

$$\begin{aligned} \bar{\delta}(t) &= \bar{\tau} - \bar{\zeta}(1 - t) \\ \bar{\zeta}(t) &= \beta_3 + \kappa \mathbb{E}(M_i \mid T_i = t) + \rho_t \sigma \sqrt{\mathbb{V}(M_i \mid T_i = t)} \\ &\quad + (\xi_3 + \mu_3)^\top \mathbb{E}(W_i \mid T_i = 1) - \xi_3^\top \mathbb{E}(W_i \mid T_i = 0) \end{aligned}$$

  where $\rho_t = \mathrm{Corr}(M_i(t, W_i(t)), \kappa_i)$ and $\sigma = \sqrt{\mathbb{V}(\kappa_i)}$ are the only unidentified quantities

- Sensitivity analysis: Examine how $\bar{\delta}(t)$ varies as a function of ρ_t

Remarks on the Proposed Sensitivity Analysis

- Interpretation of ρ_t difficult
  $\longrightarrow$ Set $\rho_t \in [-1, 1]$ and examine sharp bounds on $\bar{\delta}(t)$ as functions of σ

- Point identification under the homogeneous interaction assumption:

$$Y_i(1, m, W_i(1)) - Y_i(0, m, W_i(0)) = B_i + Cm$$

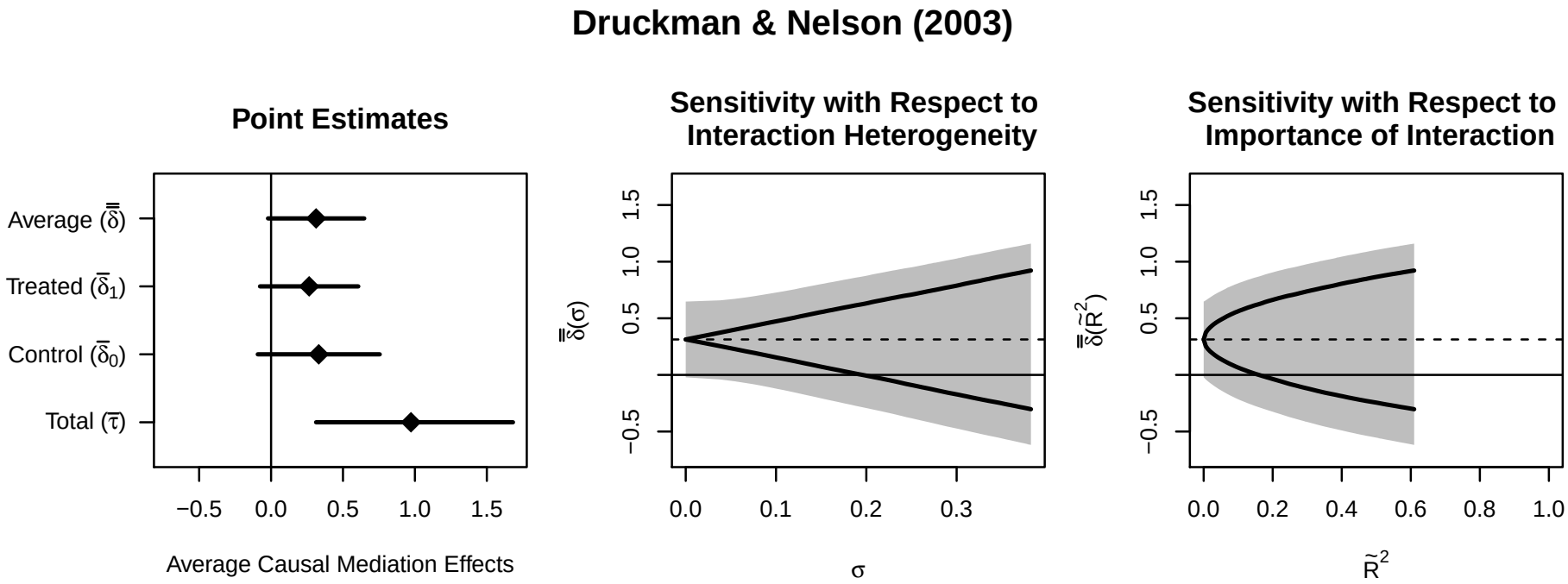
- The causal mechanism is identified as long as the degree of T–M interaction does not vary across units

- Alternative formulation using R^2 for easier interpretation:

$$R^{2*} = \frac{\mathbb{V}(\tilde{\kappa}_i T_i M_i)}{\mathbb{V}(\eta_{3i}(T_i, M_i, W_i))} \quad \text{and} \quad \widetilde{R}^2 = \frac{\mathbb{V}(\tilde{\kappa}_i T_i M_i)}{\mathbb{V}(Y_i)}$$

- How much variation in Y_i would the interaction heterogeneity have to explain for the estimate to be zero?

Reanalysis of Druckman and Nelson



- Mediation effects insignificant at 90% ($[-0.021, 0.648]$)
- Lower bound on $\bar{\delta}$ equals zero when $\sigma = 0.195$, i.e. when σ is about half as large as its largest possible value
- Effect would go away if the interaction heterogeneity explained 15.9% of the total variance of the outcome variable

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 questions and suggestions:

kimai@princeton.edu
tingley@gov.harvard.edu