Heterogeneous Treatment Effects and Causal Mechanisms

Jiawei Fu and Tara Slough

New York University

State of the field

Credibility revolution $\longrightarrow$ use of research designs that facilitate identification and estimation of causal effects.

State of the field

Credibility revolution $\longrightarrow$ use of research designs that facilitate identification and estimation of causal effects.

... but estimated causal effects do not (alone) tell us *why* or *how*

- Ultimately a question about causal mechanisms

State of the field

Credibility revolution $\longrightarrow$ use of research designs that facilitate identification and estimation of causal effects.

... but estimated causal effects do not (alone) tell us *why* or *how*

- Ultimately a question about causal mechanisms

Various approaches to evaluating mechanisms:

- Heterogeneous treatment effects (HTEs) estimated by treatment $\times$ covariate interactions is very popular

HTEs and mechanisms: a survey

We classify articles in 3 leading political science journals in 2021:

HTEs and mechanisms: a survey

We classify articles in 3 leading political science journals in 2021:

Journal	Number of: Articles	Quant. articles	Pr(Report HTE| Quant. article)	Pr(Mechanism test| Report HTE)
AJPS (65)	61	41	0.56	0.87
APSR (115)	106	75	0.53	0.90
JoP (83)	142	106	0.55	0.83
Total	309	222	0.55	0.87

HTEs and mechanisms: a survey

We classify articles in 3 leading political science journals in 2021:

Journal	Number of: Articles	Quant. articles	Pr(Report HTE| Quant. article)	Pr(Mechanism test| Report HTE)
AJPS (65)	61	41	0.56	0.87
APSR (115)	106	75	0.53	0.90
JoP (83)	142	106	0.55	0.83
Total	309	222	0.55	0.87

Takeaways:

1. Modal empirical article reports HTEs (treatment $\times$ covariate).
2. 87% of these articles use HTEs to "test mechanisms."

Known vs. under-explored problems

Usual criticism of HTEs rests on statistical issues:

- Interactions are generally underpowered.
- Multiple comparisons problems.

Known vs. under-explored problems

Usual criticism of HTEs rests on statistical issues:

- Interactions are generally underpowered.
- Multiple comparisons problems.

We abstract from these problems by:

- Assuming an infinite sample.
- Looking at one covariate with specific relation to mechanisms.

Known vs. under-explored problems

Usual criticism of HTEs rests on statistical issues:

- Interactions are generally underpowered.
- Multiple comparisons problems.

We abstract from these problems by:

- Assuming an infinite sample.
- Looking at one covariate with specific relation to mechanisms.

Under-explored problem:

Under what conditions do HTEs provide evidence of mechanism activation?

Outline

Motivating example: Exogenous shocks and voting behavior

- Based on model by Ashworth et al. (2018).
- Shows that HTEs can emerge when associated mechanism is inert.

Outline

Motivating example: Exogenous shocks and voting behavior

Framework: We develop a framework to connect causal mechanisms to HTE with respect to covariates.

- Builds from causal mediation framework (Imai et al., 2010)
- New concepts, assumptions needed for the HTE setting.

Outline

Motivating example: Exogenous shocks and voting behavior

Framework: We develop a framework to connect causal mechanisms to HTE with respect to covariates.

Results: What do we learn about a mechanism from the existence or non-existence of HTE?

- For outcomes that are *directly affected* by mechanism, HTE indicative of mechanism activation under assumptions.
- For outcomes that are *indirectly affected* by mechanism, HTE are not necessarily indicative of mechanism activation.

Outline

Motivating example: Exogenous shocks and voting behavior

Framework: We develop a framework to connect causal mechanisms to HTE with respect to covariates.

Results: What do we learn about a mechanism from the existence or non-existence of HTE?

Discussion: Using these results to inform research design.

Motivating Example

Exogenous shocks and voting

Natural experiment on effect of an exogenous shock, ω, on voter behavior:

- A natural disaster (e.g., Healy and Malhotra, 2010; Huber et al., 2012)
- An economic crisis (e.g., Wolfers, 2002)
- A pandemic (e.g., Achen and Bartels, 2004; Baccini et al., 2021)

Exogenous shocks and voting

Natural experiment on effect of an exogenous shock, ω, on voter behavior:

- A natural disaster (e.g., Healy and Malhotra, 2010; Huber et al., 2012)
- An economic crisis (e.g., Wolfers, 2002)
- A pandemic (e.g., Achen and Bartels, 2004; Baccini et al., 2021)

Example: Ashworth, Bueno de Mesquita, Friedenberg (2018):

- Assume our adaption of model is true.
- Suppose we could measure (some) model parameters directly.
  - $\rightarrow$ Characterize causal estimands in terms of these parameters.
- Ask: Can HTEs provide evidence of voter learning mechanism?

Model set-up

Incumbent at time of shock is of type $\theta \in \{\underline{\theta}, \overline{\theta}\}$, where $\overline{\theta} > \underline{\theta}$.

Model set-up

Incumbent at time of shock is of type $\theta \in \{\underline{\theta}, \overline{\theta}\}$, where $\overline{\theta} > \underline{\theta}$.

Voters do not observe θ but use governance outcome, g, to update:

$$g = f(\theta, \omega) + \varepsilon.$$

- Higher values of ω correspond to a more adverse shock
- ε is idiosyncratic shock drawn from symmetric, differentiable density, ϕ, that satisfies monotone likelihood ratio property relative to g.

Voter utility

Each voter's utility from a vote for politician, $p \in \{I, C\}$ is given by:

$$u_i^p = \theta^p + v_i \mathbb{I}(p = I)$$

Voter utility

Each voter's utility from a vote for politician, $p \in \{I, C\}$ is given by:

$$u_i^p = \theta^p + v_i \mathbb{I}(p = I)$$

Variation in the population of voters:

- $v_i \sim U(-1, 1)$ is a valence shock for the incumbent.
- Heterogeneous priors about the incumbent: $\pi_i^I \sim f_\pi$ with support on $(0, 1)$.
- Common prior about the challenger: $\pi^C \in (0, 1)$.

Sequence, voter behavior

Sequence:

1. Nature reveals shock, ω, and voters observe both ω and g.
2. Voters update their beliefs about the incumbent's type.
3. Voters vote for either the incumbent or the challenger.

Sequence, voter behavior

Sequence:

1. Nature reveals shock, ω, and voters observe both ω and g.
2. Voters update their beliefs about the incumbent's type.
3. Voters vote for either the incumbent or the challenger.

Voters' posteriors:

$$\beta(\overline{\theta}|\pi_i^I,\ \omega) = \frac{1}{1 + \frac{1-\pi_i^I}{\pi_i^I}\frac{\phi(g-f(\underline{\theta},\ \omega))}{\phi(g-f(\overline{\theta},\ \omega))}}$$

Sequence, voter behavior

Sequence:

1. Nature reveals shock, ω, and voters observe both ω and g.
2. Voters update their beliefs about the incumbent's type.
3. Voters vote for either the incumbent or the challenger.

Voters' posteriors:

$$\beta(\overline{\theta}|\pi_i^I,\ \omega) = \frac{1}{1 + \frac{1-\pi_i^I}{\pi_i^I} \frac{\phi(g - f(\underline{\theta}, \omega))}{\phi(g - f(\overline{\theta}, \omega))}}$$

A voter will vote for the incumbent if:

$$\underbrace{\beta(\overline{\theta}|\pi_i^I, \omega) + v_i}_{E[u_i^I]} \geqslant \underbrace{\pi^C}_{E[u_i^C]}$$

From theory to empirics

Treatment: Binary exposure to the shock $\omega \in \{\omega', \omega''\}$

From theory to empirics

Treatment: Binary exposure to the shock $\omega \in \{\omega', \omega''\}$

Outcomes: For the sake of exposition, consider two outcomes:

- Voter utility from the incumbent:

$$y_{1i} \equiv \beta(\overline{\theta}|\pi_i^I, \omega) + v_i$$

- Voter votes for the incumbent:

$$y_{2i} \equiv \mathbb{I}[\beta(\overline{\theta}|\pi_i^I, \omega) + v_i \geqslant \pi^C]$$

From theory to empirics

Treatment: Binary exposure to the shock $\omega \in \{\omega', \omega''\}$

Outcomes: For the sake of exposition, consider two outcomes:

- Voter utility from the incumbent:

$$y_{1i} \equiv \beta(\overline{\theta}|\pi_i^I, \omega) + v_i$$

- Voter votes for the incumbent:

$$y_{2i} \equiv \mathbb{I}[\beta(\overline{\theta}|\pi_i^I, \omega) + v_i \geqslant \pi^C]$$

Mechanism: Voter learning, not valence, since ω enters through voter's posterior.

Aside: DAG representation of interactions

Representation of "interaction" effects in DAGs is not standard (Nilsson et al., 2020)

- We need to be a bit more precise in this talk

Aside: DAG representation of interactions

Representation of "interaction" effects in DAGs is not standard (Nilsson et al., 2020)

- We need to be a bit more precise in this talk

The standard view

$$\begin{tikzcd} & X \arrow[d] \\ Z \arrow[r] & Y \end{tikzcd}$$

$$\frac{\partial Y}{\partial Z} \neq 0$$

$$\frac{\partial Y}{\partial X} \neq 0$$

$$\frac{\partial^2 Y}{\partial Z \partial X} \stackrel{?}{=} 0$$

Our notation

$$\begin{tikzcd} & X \arrow[d] \\ Z \arrow[r] & Y \end{tikzcd}$$

$$\frac{\partial Y}{\partial Z} \neq 0$$

$$\frac{\partial Y}{\partial X} \neq 0$$

$$\frac{\partial^2 Y}{\partial Z \partial X} = 0$$

$$\begin{tikzcd} & X \arrow[d] & \\ Z \arrow[rr] & {} & Y \end{tikzcd}$$

$$\frac{\partial Y}{\partial Z} \neq 0$$

$$\frac{\partial Y}{\partial X} \neq 0$$

$$\frac{\partial^2 Y}{\partial Z \partial X} \neq 0$$

The empiricist's question

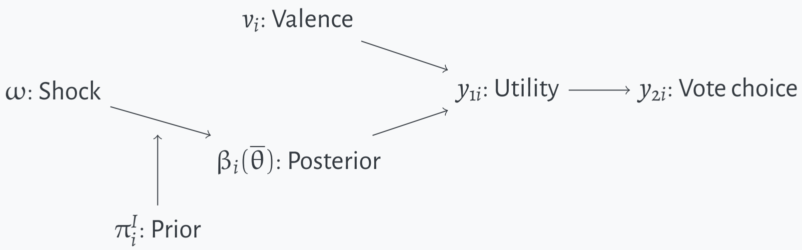
Is the mechanism:

- Voter learning about I's type? $\longleftarrow$ the mechanism
- Amplication of I's valence? $\longleftarrow$ NOT the mechanism

The empiricist's question

Is the mechanism:

- Voter learning about I's type? $\longleftarrow$ the mechanism
- Amplication of I's valence? $\longleftarrow$ NOT the mechanism



Defining HTEs

To evaluate mechanisms, the empiricist will estimate CATEs at different for different levels of the (candidate) moderators: $x \in \{\pi_i^I, v_i\}$:

$$CATE(x') = E[y(\omega'') - y(\omega')|x = x']$$

Defining HTEs

To evaluate mechanisms, the empiricist will estimate CATEs at different for different levels of the (candidate) moderators: $x \in \{\pi_i^I, v_i\}$:

$$CATE(x') = E[y(\omega'') - y(\omega')|x = x']$$

There exist HTEs in x if, for any $x' \neq x'' \in x$:

$$CATE(x'') - CATE(x') \neq 0.$$

Defining HTEs

To evaluate mechanisms, the empiricist will estimate CATEs at different for different levels of the (candidate) moderators: $x \in \{\pi_i^I, v_i\}$:

$$CATE(x') = E[y(\omega'') - y(\omega')|x = x']$$

There exist HTEs in x if, for any $x' \neq x'' \in x$:

$$CATE(x'') - CATE(x') \neq 0.$$

We will evaluate the presence of HTE for:

- Outcomes: $y \in \{$Voter utility for I, Vote for $I\}$
- Potential moderators: $x \in \{$Prior belief about I, Valence$\}$

HTEs and mechanisms (results)

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i)	Mechanism HTE $CATE(\pi) \neq CATE(\pi')$	
x_2: Valence (v_i)		

HTEs and mechanisms (results)

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i)	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$	$CATE(\pi') \neq CATE(\pi'')$
x_2: Valence (v_i)	Not a mechanism No HTE $CATE(v) = CATE(v')$	$CATE(\pi') \neq CATE(\pi'')$

HTEs and mechanisms (results)

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i^I)	Mechanism HTE $CATE(\pi) \neq CATE(\pi')$	Mechanism HTE $CATE(\pi) \neq CATE(\pi')$
x_2: Valence (v_i)	Not a mechanism No HTE $CATE(v) = CATE(v')$	$CATE(\pi') \neq CATE(\pi'')$

HTEs and mechanisms (results)

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i^I)	Mechanism HTE $CATE(\pi) \neq CATE(\pi')$	Mechanism HTE $CATE(\pi) \neq CATE(\pi')$
x_2: Valence (v_i)	Not a mechanism No HTE $CATE(v) = CATE(v')$	Not a mechanism HTE $CATE(v) \neq CATE(v')$

HTEs and mechanisms (results)

	y_1: Voter utility	y_2: Vote choice
x_1: Prior (π_i^I)	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$	Mechanism HTE $CATE(\pi') \neq CATE(\pi'')$
x_2: Valence (v_i)	Not a mechanism No HTE $CATE(v) = CATE(v')$	Not a mechanism HTE $CATE(v') \neq CATE(v')$

HTE are not necessarily indicative of mechanism activation.

- To what extent is this general?

Framework

Three main components

A treatment, Z.

Three main components

A treatment, Z.

An outcome, Y.

- Starting assumption: we measure an outcome that is *directly affected* by a mechanism. Examples:
  - $\longrightarrow$ Utility (not choice)
  - $\longrightarrow$ Latent attitudes (not Likert responses)

Three main components

A treatment, Z.

An outcome, Y.

- Starting assumption: we measure an outcome that is *directly affected* by a mechanism. Examples:
  - $\longrightarrow$ Utility (not choice)
  - $\longrightarrow$ Latent attitudes (not Likert responses)

A set of pre-treatment covariates, X.

Causal Effects

Mediators as mechanism representations.

Causal Effects

Mediators as mechanism representations.

Several causal effects typically described wrt causal mediation.

- Total effect (TE) of Z on Y.
- Indirect effect (IE_j) of Z on Y through mechanism $j \in \{1, \ldots, J\}$.
- Direct (unmediated) effect (DE) of Z on Y.

Causal Effects

Mediators as mechanism representations.

Several causal effects typically described wrt causal mediation.

- Total effect (TE) of Z on Y.
- Indirect effect (IE_j) of Z on Y through mechanism $j \in \{1, \ldots, J\}$.
- Direct (unmediated) effect (DE) of Z on Y.

At the individual/unit level:

$$TE = DE + \sum_{j=1}^{J} IE_j$$

Causal Effects

Mediators as mechanism representations.

Several causal effects typically described wrt causal mediation.

- Total effect (TE) of Z on Y.
- Indirect effect (IE_j) of Z on Y through mechanism $j \in \{1, \dots, J\}$.
- Direct (unmediated) effect (DE) of Z on Y.

At the individual/unit level:

$$TE = DE + \sum_{j=1}^{J} IE_j$$

If a mechanism j is activated or present (for any unit), then there exists some unit for which $IE_j \neq 0$.

Estimands

Average treatment effect (ATE):

$$\begin{aligned} ATE &= E_X[Y(z) - Y(z')] \\ &= ADE(z, z'; X) + \sum_{j=1}^{J} AIE(z, z'; X) \end{aligned}$$

Estimands

Average treatment effect (ATE):

$$\begin{aligned} ATE &= E_X[Y(z) - Y(z')] \\ &= ADE(z, z'; X) + \sum_{j=1}^{J} AIE(z, z'; X) \end{aligned}$$

Conditional average treatment effects (CATE): Consider pre-treatment covariate $X_k \in X$. The CATE with respect to $X_k = x$ is:

$$CATE(X_k = x) = E_{X_{\neg k}}[Y(z) - Y(z')|X_k = x]$$

Estimands

Average treatment effect (ATE):

$$
\begin{aligned}
ATE &= E_X[Y(z) - Y(z')] \\
&= ADE(z, z'; X) + \sum_{j=1}^{J} AIE(z, z'; X)
\end{aligned}
$$

Conditional average treatment effects (CATE): Consider pre-treatment covariate $X_k \in X$. The CATE with respect to $X_k = x$ is:

$$CATE(X_k = x) = E_{X_{\neg k}}[Y(z) - Y(z')|X_k = x]$$

Heterogeneous Treatment Effects (HTEs): HTEs exist with respect to pre-treatment covariate $X_k \in X$ iff:

$$\text{CATE}(\text{X}_k = x) \neq CATE(X_k = x')$$

for some $x \neq x' \in X_k$.

Reformulating the question

Original statement:

Under what conditions do HTEs provide evidence of mechanism activation?

Reformulating the question

Original statement:

Under what conditions do HTEs provide evidence of mechanism activation?

More precise version:

Under what conditions are HTEs with respect to X_k sufficient to show that there there exists some unit for which $IE_j \neq 0$?

Relationship to mediation

Mediation is advocated as a method for quantitative evaluation of mechanisms.

Relationship to mediation

Mediation is advocated as a method for quantitative evaluation of mechanisms.

Mediation:

- Requires mediators to be measurable and measured.
- Assumes sequential ignorability.
- Seeks to estimate or bound IE_j and DE directly.

Relationship to mediation

Mediation is advocated as a method for quantitative evaluation of mechanisms.

Mediation:

- Requires mediators to be measurable and measured.
- Assumes sequential ignorability.
- Seeks to estimate or bound IE_j and DE directly.

Use of HTE:

- Does not require mediators to be measurable, But we need specific measured covariates.
- Invokes a set of exclusion assumptions.
- Seeks to demonstrate that $IE_j \neq 0$ for some unit.

HTEs and Mechanisms:
Directly Affected Outcomes

Concept: Causal Indicator Variable (CIV)

Definition (Causal Indicator Variable)

Pre-treatment variable X_k is a causal indicator variable (CIV) for mechanism 1 if for some $x, x' \in X^k$, $IE_1(X_k = x) \neq IE_1(X_k = x')$.

Concept: Causal Indicator Variable (CIV)

Definition (Causal Indicator Variable)

Pre-treatment variable X_k *is a causal indicator variable (CIV) for mechanism 1 if for some* $x, x' \in X^k$, $IE_1(X_k = x) \neq IE_1(X_k = x')$.

X^{CIV} is the (possibly empty) set of covariates that satisfy definition.

Concept: Causal Indicator Variable (CIV)

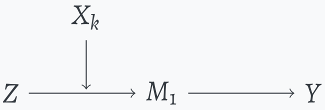
Definition (Causal Indicator Variable)

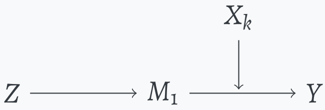
Pre-treatment variable X_k is a causal indicator variable (CIV) for mechanism 1 if for some $x, x' \in X^k$, $IE_1(X_k = x) \neq IE_1(X_k = x')$.

X^{CIV} is the (possibly empty) set of covariates that satisfy definition.

Two possibilities:

- $X_k \in X^{CIV}$ moderates the effect of treatment (Z) on mediator (M_j).
- $X_k \in X^{CIV}$ moderates the effect of the mediator (M_j) on outcome (Y).



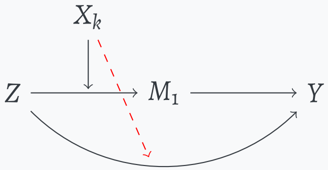


Exclusion assumption I

Assumption (Exclusion I)

Given $z, z' \in Z$ and $x, x' \in X_k$, X_k is excluded to the direct effect such that $ADE(z, z'; X_k = x) = ADE(z, z'; X = x')$.

- Direct effect of Z on Y cannot depend on X_k.

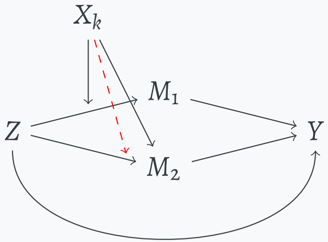


Exclusion assumption II

Assumption (Exclusion II)

Given $z, z' \in Z$ and $x, x' \in X_k$, X_k is excluded to the indirect effect of any other mechanism, $j' \neq j$, $IE_{j'}$, if: $AIE_{j'}(x) = AIE_{j'}(x')$.

- In other words, X_k is not a CIV for M_2.



HTEs as a test of mechanisms (#1 of 2)

Proposition

Suppose that Y is directly affected by mechanism j and Assumptions 1 and 2 hold with respect to X_k. If HTEs exist with respect to X_k, then $X_k \in \boldsymbol{X}^{CIV}$ for mechanism j.

HTEs as a test of mechanisms (#1 of 2)

Proposition

Suppose that Y is directly affected by mechanism j and Assumptions 1 and 2 hold with respect to X_k. If HTEs exist with respect to X_k, then $X_k \in \boldsymbol{X}^{CIV}$ for mechanism j.

Implication: By definition of CIV, HTEs imply that $IE_1(X_k = x') \neq IE_1(X_k = x'')$ for some $x', x'' \in X_k$, which indicates that mechanism j is active.

HTEs as a test of mechanisms (#1 of 2)

Proposition

Suppose that Y is directly affected by mechanism j and Assumptions 1 and 2 hold with respect to X_k. If HTEs exist with respect to X_k, then $X_k \in \boldsymbol{X}^{CIV}$ for mechanism j.

Implication: By definition of CIV, HTEs imply that $IE_1(X_k = x') \neq IE_1(X_k = x'')$ for some $x', x'' \in X_k$, which indicates that mechanism j is active.

The usual logic for HTE, but note the assumptions.

HTEs as a test of mechanisms (#2 of 2)

Proposition

Suppose that Y is directly affected by mechanism j and Assumptions 1 and 2 hold. If no HTEs exist with respect to X_k, at least one of the following must be true:

1. $X_k \notin \boldsymbol{X}^{\text{CIV}}$ *for mechanism j.*
2. *No CIV exists.*

HTEs as a test of mechanisms (#2 of 2)

Proposition

Suppose that Y is directly affected by mechanism j and Assumptions 1 and 2 hold. If no HTEs exist with respect to X_k, at least one of the following must be true:

1. $X_k \notin \boldsymbol{X}^{CIV}$ *for mechanism j.*
2. *No CIV exists.*

Implications: We don't learn about mechanism activation from lack of HTE.

- #1 implies our model is wrong. Mechanism j could be active or inert.
- An active mechanism may not have any CIV. An inert mechanism has no CIV by definition.

HTEs as a test of mechanisms (#2 of 2)

Proposition

Suppose that Y is directly affected by mechanism j and Assumptions 1 and 2 hold. If no HTEs exist with respect to X_k, at least one of the following must be true:

1. $X_k \notin \boldsymbol{X}^{CIV}$ *for mechanism j.*
2. *No CIV exists.*

Implications: We don't learn about mechanism activation from lack of HTE.

- #1 implies our model is wrong. Mechanism j could be active or inert.
- An active mechanism may not have any CIV. An inert mechanism has no CIV by definition.

HTEs as a test of mechanisms (#2 of 2)

Proposition

Suppose that Y is directly affected by mechanism j and Assumptions 1 and 2 hold. If no HTEs exist with respect to X_k, at least one of the following must be true:

1. $X_k \notin \boldsymbol{X}^{CIV}$ *for mechanism* j.
2. *No CIV exists.*

Implications: We don't learn about mechanism activation from lack of HTE.

- #1 implies our model is wrong. Mechanism j could be active or inert.
- An active mechanism may not have any CIV. An inert mechanism has no CIV by definition.

Absence of HTE does not "rule out" a mechanism.

HTE and Mechanisms: Indirectly Affected Outcomes

Why should we care?

Many attitudinal, behavioral outcomes are indirectly affected by mechanisms.

- Example: decisions based on utility maximization.
- Vote choice vs. utility in our motivating example

Why should we care?

Many attitudinal, behavioral outcomes are indirectly affected by mechanisms.

- Example: decisions based on utility maximization.
- Vote choice vs. utility in our motivating example

Matters whenever mechanisms act upon a latent, directly affected outcome.

- Much of the time.
- In the example, voter learning changes expected utility.

Why should we care?

Many attitudinal, behavioral outcomes are indirectly affected by mechanisms.

- Example: decisions based on utility maximization.
- Vote choice vs. utility in our motivating example

Matters whenever mechanisms act upon a latent, directly affected outcome.

- Much of the time.
- In the example, voter learning changes expected utility.

Poses challenges for the quantitative detection of mechanisms

- Through HTEs and likely other approaches.

Additional structure, concept

Suppose that Y is directly-affected outcome. We observe $L(Y)$.

- Concern emerges when $L(\cdot)$ is non-linear.
- e.g., Discrete outcomes (choices, attitudinal scales etc.)

Additional structure, concept

Suppose that Y is directly-affected outcome. We observe $L(Y)$.

- Concern emerges when $L(\cdot)$ is non-linear.
- e.g., Discrete outcomes (choices, attitudinal scales etc.)

Useful to define $\mathbf{X}^R$ as the subset of measured covariates with a non-zero effect on outcome Y. It is straightforward to see that:

$$\mathbf{X}^{CIV} \subseteq \mathbf{X}^R \subseteq \mathbf{X}$$

Additional structure, concept

Suppose that Y is directly-affected outcome. We observe $L(Y)$.

- Concern emerges when $L(\cdot)$ is non-linear.
- e.g., Discrete outcomes (choices, attitudinal scales etc.)

Useful to define $\mathbf{X}^R$ as the subset of measured covariates with a non-zero effect on outcome Y. It is straightforward to see that:

$$\mathbf{X}^{CIV} \subseteq \mathbf{X}^R \subseteq \mathbf{X}$$

In our motivating example, for the learning mechanism:

- $\mathbf{X}^{CIV} = \{\pi_i^I\}$
- $\mathbf{X}^R = \{\pi_i^I, v_i\}$

HTEs as a test of mechanisms (#1 of 2)

Proposition

Suppose that observed outcome $L(Y)$ is a non-linear mapping of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTEs exist with respect to X_k, then $X_k \in \boldsymbol{X}^R$.

HTEs as a test of mechanisms (#1 of 2)

Proposition

Suppose that observed outcome $L(Y)$ is a non-linear mapping of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTEs exist with respect to X_k, then $X_k \in \boldsymbol{X}^R$.

Intuition: Using HTE for mechanism detection relies on *additive separability* of X_k from DE and $IE_{\neg j}$ on the latent variable.

- What exclusion assumptions buy us
- But a non-linear $L(\cdot)$ does not preserve additive separability on $L(Y)$.

HTEs as a test of mechanisms (#1 of 2)

Proposition

Suppose that observed outcome $L(Y)$ is a non-linear mapping of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTEs exist with respect to X_k, then $X_k \in \mathbf{X}^R$.

Intuition: Using HTE for mechanism detection relies on *additive separability* of X_k from DE and $IE_{\neg j}$ on the latent variable.

- What exclusion assumptions buy us
- But a non-linear $L(\cdot)$ does not preserve additive separability on $L(Y)$.

Implication: Two possibilities:

- $X_k \in \mathbf{X}^{CIV} \implies$ mechanism j is active.
- $X_k \notin \mathbf{X}^{CIV} \implies$ mechanism j may or may not be active.

Numerical example

Suppose that we are interested in how a mobilization treatment, $Z_i \in \{0, 1\}$, affects voters' turnout decisions. Two covariates in $\mathbf{X}^R$:

- $X_1 \sim \mathcal{N}(0, 1)$
- $X_2 \sim \text{Bernoulli}(0.5)$

Numerical example

Suppose that we are interested in how a mobilization treatment, $Z_i \in \{0, 1\}$, affects voters' turnout decisions. Two covariates in $\mathbf{X}^R$:

- $X_1 \sim \mathcal{N}(0, 1)$
- $X_2 \sim \text{Bernoulli}(0.5)$

X_1 is a CIV for the only mechanism M_1, such that:

$$M_1(Z, \mathbf{X}) = (1 + Z)X_1$$

Numerical example

Suppose that we are interested in how a mobilization treatment, $Z_i \in \{0, 1\}$, affects voters' turnout decisions. Two covariates in $\mathbf{X}^R$:

- $X_1 \sim \mathcal{N}(0, 1)$
- $X_2 \sim \text{Bernoulli}(0.5)$

X_1 is a CIV for the only mechanism M_1, such that:

$$M_1(Z, \mathbf{X}) = (1 + Z)X_1$$

Potential voters' utility from voting is given by:

$$U(Z, X) = M_1(Z, X_1) + X_2 = (1 + Z)X_1 + X_2$$

Numerical example

Suppose that we are interested in how a mobilization treatment, $Z_i \in \{0, 1\}$, affects voters' turnout decisions. Two covariates in $\mathbf{X}^R$:

- $X_1 \sim \mathcal{N}(0, 1)$
- $X_2 \sim \text{Bernoulli}(0.5)$

X_1 is a CIV for the only mechanism M_1, such that:

$$M_1(Z, \mathbf{X}) = (1 + Z)X_1$$

Potential voters' utility from voting is given by:

$$U(Z, X) = M_1(Z, X_1) + X_2 = (1 + Z)X_1 + X_2$$

Turnout—the observed outcome—is given by:

$$L(U(Z, X)) = \begin{cases} 1 & \text{if } (1 + Z)X_1 + X_2 \geqslant 0 \\ 0 & \text{else} \end{cases}$$

HTEs with respect to X_2?

Recall that X_2 is **not** a CIV for the unique mechanism, M_1.

$CATE(X_2 = 1)$ is given by:

$$\begin{aligned} CATE(X_2 = 1) &= E[L(U(Z=1, X)) - L(U(Z=0, X))|X_2 = 1] \\ &= \Pr(2X_1 + 1 > 0) - \Pr(X_1 + 1 > 0) \\ &= \Phi(-1) - \Phi(-\frac{1}{2}) \approx -0.15 \end{aligned}$$

It is straightforward to see that $CATE(X_2 = 0) = \Phi(0) - \Phi(0) = 0$.

HTEs exist with respect to X_2, but we know that $X_2 \in \mathbf{X}^R - \mathbf{X}^{CIV}$.

HTEs as a test of mechanisms (#2 of 2)

Proposition

Suppose that observed outcome $L(Y)$ is a non-linear mapping of directly-affected outcome Y and Assumptions 1 and 2 hold. If HTEs do not exist with respect to X_k, then $X_k \in \mathbf{X}$.

Implication: This is vacuous! Obviously measured covariate X_k is in the set of measured covariates $\mathbf{X}$...

- Without further assumptions about distribution of Y and functional form of $L(\cdot)$ we cannot say anything from a lack of heterogeneity for an indirectly affected outcome!

Summary of results

	Outcome variable is: Directly affected	Outcome variable is: Indirectly affected
$\exists$ HTEs wrt X_k:	$X_k \in \mathbf{X}^{CIV}$	$X_k \in \mathbf{X}^{R}$
	$\implies M_j$ is active.	M_j active or inactive
$\nexists$ HTEs wrt X_k:	$X_k \notin \mathbf{X}^{CIV}$ **and/or** $\nexists$ CIV for mechanism j	$X_k \in \mathbf{X}$ (vacuous)
	M_j active or inactive	M_j active or inactive

Recommendations for Research Design

Recommendation #1: Theoretical questions

Three questions needed to support use of HTEs for mechanism detection:

1. Enumeration of set of candidate mechanisms.

Recommendation #1: Theoretical questions

Three questions needed to support use of HTEs for mechanism detection:

1. Enumeration of set of candidate mechanisms.
2. Relationship between a covariate, X_k and each candidate mechanism:
   - $\longrightarrow$ For which mechanism (j) is X_k a candidate CIV?
   - $\longrightarrow$ Is exclusion assumption plausible for every other candidate mechanism?

Recommendation #1: Theoretical questions

Three questions needed to support use of HTEs for mechanism detection:

1. Enumeration of set of candidate mechanisms.
2. Relationship between a covariate, X_k and each candidate mechanism:
   - $\longrightarrow$ For which mechanism (j) is X_k a candidate CIV?
   - $\longrightarrow$ Is exclusion assumption plausible for every other candidate mechanism?
3. Classification of mechanisms as *directly affected* or *indirectly affected*
   - $\longrightarrow$ In some theory traditions, requires more theoretical structure.
   - $\longrightarrow$ We should likely focus on HTE for some outcomes but not others.

Recommendation #2: Improving interpretation of HTEs

Absence of HTEs do not "rule out" candidate mechanisms.

- Given a candidate CIV, presence of HTEs is informative only when: exclusion assumptions hold, outcome is directly affected.
- HTEs provide *less information* than is generally asserted by their interpretation.

Recommendation #2: Improving interpretation of HTEs

Absence of HTEs do not “rule out” candidate mechanisms.

- Given a candidate CIV, presence of HTEs is informative only when: exclusion assumptions hold, outcome is directly affected.
- HTEs provide *less information* than is generally asserted by their interpretation.

Implications of low power for interactions we are less likely to *detect* HTE that do exist.

- Compounds these challenges of interpretation.

Recommendation #3: Improving research design

To use HTEs for mechanism detection, the the more measured candidate CIVs is better.

- We need multiple candidate CIVs to satisfy exclusion assumptions...
- Benefit: mixed results (HTEs in one candidate but not another) resolve ambiguity about existence of CIVs.
  - $\longrightarrow$ For directly-affected outcomes, permits attribution of lack of HTE in one candidate to mis-specification.

Recommendation #3: Improving research design

To use HTEs for mechanism detection, the the more measured candidate CIVs is better.

- We need multiple candidate CIVs to satisfy exclusion assumptions...
- Benefit: mixed results (HTEs in one candidate but not another) resolve ambiguity about existence of CIVs.
  - $\longrightarrow$ For directly-affected outcomes, permits attribution of lack of HTE in one candidate to mis-specification.

Clearer specification of relationship between mechanisms and outcomes

- Prioritize directly-affected outcomes for mechanism detection
  - $\longrightarrow$ Design measurement instruments to elicit these outcomes.
  - $\longrightarrow$ *Maybe* an argument for latent variable models?

Recommendation #4: Adding assumptions?

Can assumption of monotonicity make HTE more informative in the context of indirectly-observed outcomes?

for all x'>x $\in X_k, CATE(x') > (<) CATE(x)$

Recommendation #4: Adding assumptions?

Can assumption of monotonicity make HTE more informative in the context of indirectly-observed outcomes?

for all x'>x $\in X_k, CATE(x') > (<) CATE(x)$

Answer: No, not in isolation. We need additional assumptions about:

- The DGP in the form of the empirical distribution of Y or the distribution of any error terms.
- The mapping $L(\cdot)$.

Conclusion

Four takeaways

1. A problem: the use of HTEs for mechanism detection is very popular but under-theorized.
2. HTEs is not a “agnostic” approach to analysis of mechanisms: requires exclusion assumptions.
3. This approach provides information about mechanisms when:
    - Exclusion assumptions hold
    - Outcome is directly affected by a given mechanism.
4. We can better learn about mechanisms HTEs by more carefully approaching these analyses.

Thank You!