# Instrumental Variables with Heterogeneous Treatment Effects

(in preparation for *The Handbook of Labor Economics*)

Magne Mogstad
University of Chicago
Statistics Norway
NBER

Alexander Torgovitsky[1]
University of Chicago

1 Research supported by National Science Foundation grant SES-1846832

### Topic of the chapter

- Instrumental variables (IV) with heterogeneous treatment effects (HTEs)

$\rightarrow$ *Unobservable* heterogeneity

- Complicates IV methods tremendously
- An enormous and sometimes contentious cross-disciplinary literature
- Featured centrally in three Nobel prizes (Heckman, Imbens, Angrist)

$\rightarrow$ Speaks to several fundamental issues in empirical methodology

## Topic of the chapter

- Instrumental variables (IV) with heterogeneous treatment effects (HTEs)

→ *Unobservable* heterogeneity

- Complicates IV methods tremendously
- An enormous and sometimes contentious cross-disciplinary literature
- Featured centrally in three Nobel prizes (Heckman, Imbens, Angrist)

→ Speaks to several fundamental issues in empirical methodology

## Thematic organization of the chapter/this talk

1. Background: why are HTEs important with IV? Key concepts
2. Reverse Engineering: Interpreting Linear Estimators
3. Forward Engineering: Estimating Target Parameters

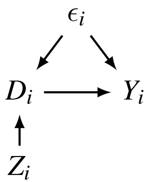
# Outline

# IV in a Nutshell



### Observed variables

- $Y_i$ outcome
- $D_i$ endogenous variable (“treatment”)
- $Z_i$ instrument




### Three assumptions in every IV context

1. **Exclusion:** $Z_i$ has no causal effect on $Y_i$
2. **Exogeneity:** $Z_i$ not associated with $\epsilon_i$
3. **Relevance:** $Z_i$ and $D_i$ are associated

### Our focus

- Allowing unobserved heterogeneity in the causal effect of $D_i$ on $Y_i$
- Taking exclusion as given, exogeneity also (w/ a caveat)

## Is this complication worth it?

- Constant effects will always be an escape back to safety
- Slightly more generally, *un*systematic HTEs (uncorrelated with $D_i$)

### Is this complication worth it?

- Constant effects will always be an escape back to safety
- Slightly more generally, *un*systematic HTEs (uncorrelated with $D_i$)

### Returns to college (Becker 1964, Griliches 1977, Card 1999)

- "Ability" — but how about heterogeneous skills/complementarity?
- College important for turning abstract reasoning skills into $$$
- Not important for turning "working with hands" skills into $$$

$\Rightarrow$ *Systematic* HTEs due to unobservables (skills)

### Is this complication worth it?

- Constant effects will always be an escape back to safety
- Slightly more generally, *un*systematic HTEs (uncorrelated with $D_i$)

### Returns to college (Becker 1964, Griliches 1977, Card 1999)

- "Ability" — but how about heterogeneous skills/complementarity?
- College important for turning abstract reasoning skills into $$$
- Not important for turning "working with hands" skills into $$$

$\Rightarrow$ *Systematic* HTEs due to unobservables (skills)

### An example of a general phenomenon

- HTEs due to nature and "production function"
- Agent chooses $D_i$ while considering effect on $Y_i$ (a common IV story!)

$\Rightarrow$ *Systematic* HTEs — unobserved heterogeneity correlated with $D_i$

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

**Two restrictions of this model (interpreted literally ...)**

1. **Linear treatment effect**: Not our focus (not restrictive if $D_i \in \{0, 1\}$)
2. **Constant treatment effect**: $\beta_1$ does not vary with $i$

$\rightarrow$ More generally, does not vary with unobservables (no HTEs)

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

**Two restrictions of this model (interpreted literally . . . )**

1. **Linear treatment effect**: Not our focus (not restrictive if $D_i \in \{0, 1\}$)
2. **Constant treatment effect**: $\beta_1$ does not vary with $i$

$\rightarrow$ More generally, does not vary with unobservables (no HTEs)

**Notation for relaxing constant treatment effects**

- **Nonseparable model:** $Y_i = g(D_i, \epsilon_i)$

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

**Two restrictions of this model (interpreted literally ...)**

1. **Linear treatment effect**: Not our focus (not restrictive if $D_i \in \{0, 1\}$)
2. **Constant treatment effect**: $\beta_1$ does not vary with $i$

$\rightarrow$ More generally, does not vary with unobservables (no HTEs)

**Notation for relaxing constant treatment effects**

- **Nonseparable model:** $Y_i = g(D_i, \epsilon_i)$
- **Potential outcomes:** $Y_i(d)$, with $Y_i = Y_i(D_i)$

$$\text{e.g.} \qquad Y_i = (1 - D_i)Y_i(0) + D_i Y_i(1) \qquad \text{if } D_i \in \{0, 1\}$$

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

**Two restrictions of this model (interpreted literally ...)**

1. **Linear treatment effect**: Not our focus (not restrictive if $D_i \in \{0, 1\}$)
2. **Constant treatment effect**: $\beta_1$ does not vary with $i$

$\rightarrow$ More generally, does not vary with unobservables (no HTEs)

**Notation for relaxing constant treatment effects**

- **Nonseparable model:** $Y_i = g(D_i, \epsilon_i)$
- **Potential outcomes:** $Y_i(d)$, with $Y_i = Y_i(D_i)$

$$\text{e.g.} \qquad Y_i = (1 - D_i)Y_i(0) + D_i Y_i(1) \qquad \text{if } D_i \in \{0, 1\}$$

- Potential outcomes notation more popular (but it's just notation)

### Why do HTEs complicate IV?

- It matters "who" takes treatment
- Modeling selection a way to organize (and restrict) this relationship
- Consider $D_i \in \{0, 1\}$ for now

### Why do HTEs complicate IV?

- It matters “who” takes treatment
- Modeling selection a way to organize (and restrict) this relationship
- Consider $D_i \in \{0, 1\}$ for now

### Threshold-crossing model (e.g. Heckman 1979)

$$D_i = \mathbb{1}[V_i \leq \gamma Z_i]$$

### Why do HTEs complicate IV?

- It matters "who" takes treatment
- Modeling selection a way to organize (and restrict) this relationship
- Consider $D_i \in \{0, 1\}$ for now

### Threshold-crossing model (e.g. Heckman 1979)

$$D_i = \mathbb{1}[V_i \leq \gamma Z_i]$$

### Potential treatments with monotonicity (Imbens & Angrist 1994)

$$D_i = (1 - Z_i)D_i(0) + Z_i D_i(1) \quad \text{and} \quad \underbrace{\mathbb{P}[D_i(1) \geq D_i(0)] = 1}_{\text{the monotonicity condition}}$$

always-takers, never-takers, compliers — no defiers

### Why do HTEs complicate IV?

- It matters “who” takes treatment
- Modeling selection a way to organize (and restrict) this relationship
- Consider $D_i \in \{0, 1\}$ for now

### Threshold-crossing model (e.g. Heckman 1979)

$$D_i = \mathbb{1}[V_i \leq \gamma Z_i]$$

### Potential treatments with monotonicity (Imbens & Angrist 1994)

$$D_i = (1 - Z_i)D_i(0) + Z_i D_i(1) \quad \text{and} \quad \underbrace{\mathbb{P}[D_i(1) \geq D_i(0)] = 1}_{\text{the monotonicity condition}}$$

always-takers, never-takers, compliers — no defiers

**Different notation for *the same* model (Vytlacil, 2002)**

### Weak exogeneity

- $(Y_i(0), Y_i(1))$ independent of $Z_i$ given covariates
- Nonparametric analog of $\mathbb{E}[\epsilon_i Z_i] = 0$

## Weak exogeneity

- $(Y_i(0), Y_i(1))$ independent of $Z_i$ given covariates
- Nonparametric analog of $\mathbb{E}[\epsilon_i Z_i] = 0$

## Strong exogeneity

- $(Y_i(0), Y_i(1), D_i(0), D_i(1))$ independent of $Z_i$ given covariates
- Equivalently, $(Y_i(0), Y_i(1), V_i)$ independent of $Z_i$ given covariates

# Weak vs. Strong Exogeneity



## Weak exogeneity

- $(Y_i(0), Y_i(1))$ independent of $Z_i$ given covariates
- Nonparametric analog of $\mathbb{E}[\epsilon_i Z_i] = 0$

## Strong exogeneity

- $(Y_i(0), Y_i(1), D_i(0), D_i(1))$ independent of $Z_i$ given covariates
- Equivalently, $(Y_i(0), Y_i(1), V_i)$ independent of $Z_i$ given covariates

## Implications of strong exogeneity

- Can identify the causal effect of the instrument on treatment
- Selection model is "causal" vs. statistical first stage
- Different correlated instruments cannot be considered in isolation

$\rightarrow$ $Z_{i2}$ is part of $V_i$ (or $D_i(0), D_i(1)$) if not controlled for

- e.g. tuition and distance in returns to college literature

### HTEs require confronting a new question

- *Who* do we want to estimate treatment effects for?
- Not a question we had to ask in the classical model (one effect: $\beta$)
- Choice will (naturally) involve trade-offs — what should guide them?

## HTEs require confronting a new question

- *Who* do we want to estimate treatment effects for?
- Not a question we had to ask in the classical model (one effect: $\beta$)
- Choice will (naturally) involve trade-offs — what should guide them?

**Target parameter:** the object we are trying to estimate

# Target Parameters

### HTEs require confronting a new question

- *Who* do we want to estimate treatment effects for?
- Not a question we had to ask in the classical model (one effect: $\beta$)
- Choice will (naturally) involve trade-offs — what should guide them?

**Target parameter:** the object we are trying to estimate

### Why are we attempting causal inference to begin with?

- Policy — trying to inform a decision

$\rightarrow$ Usually provides a clear target parameter (Heckman & Vytlacil 2005)

# Target Parameters



## HTEs require confronting a new question

- *Who* do we want to estimate treatment effects for?
- Not a question we had to ask in the classical model (one effect: $\beta$)
- Choice will (naturally) involve trade-offs — what should guide them?

**Target parameter:** the object we are trying to estimate

## Why are we attempting causal inference to begin with?

- Policy — trying to inform a decision

$\rightarrow$ Usually provides a clear target parameter (Heckman & Vytlacil 2005)

- "Science" — knowledge for the sake of knowledge (?)
- Less guidance on the appropriate target parameter
- Easy to interpret? Generalizable? Difficulty to identify/estimate?

**Reverse engineering: interpreting linear estimators**

- Start with a commonly-used estimator (linear IV/TSLS)
- Determine assumptions under which it estimates something interesting
- “Reverse” because it starts with the tool

$\rightarrow$ The literature on local average treatment effects (LATEs)

### Reverse engineering: interpreting linear estimators

- Start with a commonly-used estimator (linear IV/TSLS)
- Determine assumptions under which it estimates something interesting
- "Reverse" because it starts with the tool

$\rightarrow$ The literature on local average treatment effects (LATEs)

### Forward engineering: estimating target parameters

- Start with a target parameter
- Derive an estimator of it under some assumptions
- "Forward" because it starts with the problem

$\rightarrow$ The literature on selection corrections/control functions

**Reverse engineering requires a new definition**

- **Estim*ator*:** a procedure mapping data into a number
- **Estim*and*:** what an estimator is consistent for (what it estimates)

## Reverse engineering requires a new definition

- **Estim*ator*:** a procedure mapping data into a number
- **Estim*and*:** what an estimator is consistent for (what it estimates)

## What's the thought process?

- Classical linear IV model **misspecified** with HTEs (and/or nonlinearity)
- Maybe estimand has a useful interpretation robust to misspecification?

## Reverse engineering requires a new definition

- **Estim*ator*:** a procedure mapping data into a number
- **Estim*and*:** what an estimator is consistent for (what it estimates)

## What's the thought process?

- Classical linear IV model **misspecified** with HTEs (and/or nonlinearity)
- Maybe estimand has a useful interpretation robust to misspecification?

## Useful interpretation?

- Usually: convex-weighted average of subgroup treatment effects:

$$\text{the estimand} = \sum_g \overbrace{\omega(g)}^{\text{non-negative weights that sum to one}} \underbrace{\mathbb{E}[Y_i(1) - Y_i(0) | G_i = g]}_{\text{subgroup treatment effects}}$$

- **Weak causality:** all effects positive $\Rightarrow$ estimand positive

### Imbens & Angrist (1994) local average treatment effect (LATE)

Given monotonicity and strong exogeneity,

$$\underbrace{\frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]}}_{\text{Wald estimand}} = \underbrace{\mathbb{E}[Y_i(1)-Y_i(0)|\overbrace{D_i(0)=0,D_i(1)=1}^{\text{subpopulation of compliers } (G_i=(0,1))}]}_{\text{average treatment effect for compliers (LATE)}}$$

### Imbens & Angrist (1994) local average treatment effect (LATE)

Given monotonicity and strong exogeneity,

$$\underbrace{\frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]}}_{\text{Wald estimand}} = \underbrace{\mathbb{E}[Y_i(1)-Y_i(0)|\overbrace{D_i(0)=0, D_i(1)=1}^{\text{subpopulation of compliers } (G_i=(0,1))}]}_{\text{average treatment effect for compliers (LATE)}}$$

### Misspecification-robust interpretation of linear IV

$$\text{LATE} = \overbrace{\underbrace{\frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]}}_{\text{Wald estimand}} = \underbrace{\frac{\mathbb{C}[Y_i,Z_i]}{\mathbb{C}[D_i,Z_i]}}_{\text{simple linear IV estimand}}}^{\text{because } Z_i \text{ is binary (and } only \text{ in that case)}}$$

- The "simple" linear IV estimand instruments $[1, D_i]'$ with $[1, Z_i]'$

$\rightarrow$ e.g. `ivregress 2sls y (d = z)` coefficient on `d`

### Departures from the baseline case

- **Instrument:** multivalued — no longer a simple binary contrast

  *Weighted average of LATEs with weights that depend on the distribution of Z*

- **Assumptions:** failure of monotonicity

  *e.g. unordered instruments like judges (Frandsen et al 2023), multiple instruments (Mogstad et al 2021), or just because it's a questionable assumption (Angrist et al 1996, Angrist & Evans 1998)*

- **Treatment:** multivalued — ordered or unordered

  *Multivalued ordered extends nicely; unordered case is complicated*

- **Covariates:** controlling for them (or interacting them)

  *Parametric assumptions become necessary for weakly causal interpretation (Blandhol et al 2022)*

## Departures from the baseline case

- **Instrument:** multivalued — no longer a simple binary contrast

  *Weighted average of LATEs with weights that depend on the distribution of Z*

- **Assumptions:** failure of monotonicity

  *e.g. unordered instruments like judges (Frandsen et al 2023), multiple instruments (Mogstad et al 2021), or just because it's a questionable assumption (Angrist et al 1996, Angrist & Evans 1998)*

- **Treatment:** multivalued — ordered or unordered

  *Multivalued ordered extends nicely; unordered case is complicated*

- **Covariates:** controlling for them (or interacting them)

  *Parametric assumptions become necessary for weakly causal interpretation (Blandhol et al 2022)*

## All of these cases caveat the LATE interpretation of linear IV

- Might lead to multiple possible choices of a reasonable estimator
- Structure of estimand becomes complicated, hard to transfer
- Additional assumptions may be needed for a "good" interpretation

### Departures from the baseline case

- **Covariates:** controlling for them (or interacting them)

  *Parametric assumptions become necessary for weakly causal interpretation (Blandhol et al 2022)*

### All of these cases caveat the LATE interpretation of linear IV

- Might lead to multiple possible choices of a reasonable estimator
- Structure of estimand becomes complicated, hard to transfer
- Additional assumptions may be needed for a "good" interpretation

**Two roles for covariates**

1. Support exogeneity of the instrument
2. Reduce residual variation and help tighten inference (standard errors)

### Two roles for covariates

1. Support exogeneity of the instrument
2. Reduce residual variation and help tighten inference (standard errors)

### Nonparametric conditioning

- Sample selection (e.g. married women 21–35, only married once, . . . )
- Doesn't create any conceptual complications
- But runs into the curse of dimensionality quickly

### Two roles for covariates

1. Support exogeneity of the instrument
2. Reduce residual variation and help tighten inference (standard errors)

### Nonparametric conditioning

- Sample selection (e.g. married women 21–35, only married once, . . . )
- Doesn't create any conceptual complications
- But runs into the curse of dimensionality quickly

### Linearly controlling for covariates

- The usual way of implementing fine-grained "conditioning"
- Changes the estimand in a non-obvious way

  (Linear regression is both beautiful and complicated . . . )

### The linear IV estimand

Control for a vector of covariates $X_i$:

$$\frac{\mathbb{E}[Y_i \tilde{Z}_i]}{\mathbb{E}[D_i \tilde{Z}_i]} \quad \text{where} \quad \tilde{Z}_i \equiv Z_i - \underbrace{X_i' \overbrace{\mathbb{E}[X_i X_i']^{-1}\,\mathbb{E}[X_i Z_i]}^{\text{coefficients from regressing } Z_i \text{ onto } X_i\ (\delta)}}_{\text{population fitted values from linear regression of } Z_i \text{ onto } X_i} \equiv Z_i - X_i'\delta.$$

## The linear IV estimand

Control for a vector of covariates $X_i$:

$$\frac{\mathbb{E}[Y_i\tilde{Z}_i]}{\mathbb{E}[D_i\tilde{Z}_i]} \quad \text{where} \quad \tilde{Z}_i \equiv Z_i - \underbrace{X_i' \overbrace{\mathbb{E}[X_iX_i']^{-1}\,\mathbb{E}[X_iZ_i]}^{\text{coefficients from regressing } Z_i \text{ onto } X_i\ (\delta)}}_{\text{population fitted values from linear regression of } Z_i \text{ onto } X_i} \equiv Z_i - X_i'\delta.$$

## Both $Z_i$ and $X_i$ variation gets used

$$\overbrace{\mathbb{E}[Y_i\tilde{Z}_i]}^{\text{numerator of IV estimand}} = \mathbb{E}\left[\mathbb{E}[Y_i\tilde{Z}_i|X_i]\right] = \overbrace{\mathbb{E}\left[\mathbb{C}[Y_i, Z_i|X_i]\right]}^{\text{variation in } Y_i \text{ caused by } Z_i} + \overbrace{\mathbb{E}\left[Y_i\,\mathbb{E}[\tilde{Z}_i|X_i]\right]}^{\text{covariation between } Y_i \text{ and } X_i}$$

- The first term is what we intuitively want from an IV estimand
- The second term is bad variation ...

### The problematic second term again

$$\mathbb{E}\overbrace{\left[Y_i \mathbb{E}[\tilde{Z}_i | X_i]\right]}^{\text{covariation between } Y_i \text{ and } X_i}$$

# Level-Dependence of Linear IV with Covariates



## The problematic second term again

$$\mathbb{E}\left[\overbrace{Y_i \mathbb{E}[\tilde{Z}_i | X_i]}^{\text{covariation between } Y_i \text{ and } X_i}\right]$$

## Problematic because it introduces level-dependence

- The *levels* of $Y_i$ reflect always-takers, never-takers
- Level-dependent estimands are not weakly causal

→ Effects could all be positive, but estimand negative due to levels

## The problematic second term again

$$\mathbb{E}\left[\overbrace{Y_i \mathbb{E}[\tilde{Z}_i | X_i]}^{\text{covariation between } Y_i \text{ and } X_i}\right]$$

## Problematic because it introduces level-dependence

- The *levels* of $Y_i$ reflect always-takers, never-takers
- Level-dependent estimands are not weakly causal

$\rightarrow$ Effects could all be positive, but estimand negative due to levels

## When does level-dependence go away?

- Classical linear model with constant effects
- **Rich covariates:** $\mathbb{E}[Z_i | X_i] = \delta' X_i$ is actually linear ...

**Definition**

$$\mathbb{E}[\tilde{Z}_i|X_i = x] = \underbrace{\mathbb{E}[Z_i|X_i = x]}_{\text{must be linear}} - \delta' x = 0 \quad \text{for all } x$$

### Definition

$$\mathbb{E}[\tilde{Z}_i | X_i = x] = \underbrace{\mathbb{E}[Z_i | X_i = x]}_{\text{must be linear}} - \delta' x = 0 \quad \text{for all } x$$

### Necessity of rich covariates

- If it doesn't hold, then linear IV is not weakly causal (***necessarily***)
- Level-dependence!

## Definition

$$\mathbb{E}[\tilde{Z}_i|X_i = x] = \underbrace{\mathbb{E}[Z_i|X_i = x]}_{\text{must be linear}} - \delta' x = 0 \quad \text{for all } x$$

## Necessity of rich covariates

- If it doesn't hold, then linear IV is not weakly causal (***necessarily***)
- Level-dependence!

## Ensuring rich covariates

- Satisfied if $X_i$ and $Z_i$ are independent (e.g. experiment, fuzzy RD)

### Definition

$$\mathbb{E}[\tilde{Z}_i|X_i = x] = \underbrace{\mathbb{E}[Z_i|X_i = x]}_{\text{must be linear}} - \delta' x = 0 \quad \text{for all } x$$

### Necessity of rich covariates

- If it doesn't hold, then linear IV is not weakly causal (***necessarily***)
- Level-dependence!

### Ensuring rich covariates

- Satisfied if $X_i$ and $Z_i$ are independent (e.g. experiment, fuzzy RD)
- Satisfied with an extremely flexible specification of $X_i$

  e.g. "saturate and weight" in *Mostly Harmless*

### Definition

$$\mathbb{E}[\tilde{Z}_i | X_i = x] = \underbrace{\mathbb{E}[Z_i | X_i = x]}_{\text{must be linear}} - \delta' x = 0 \quad \text{for all } x$$

### Necessity of rich covariates

- If it doesn't hold, then linear IV is not weakly causal (***necessarily***)
- Level-dependence!

### Ensuring rich covariates

- Satisfied if $X_i$ and $Z_i$ are independent (e.g. experiment, fuzzy RD)
- Satisfied with an extremely flexible specification of $X_i$

  e.g. "saturate and weight" in *Mostly Harmless*
- Otherwise it's a parametric *assumption*

$\rightarrow$ At odds with the motivation for reverse-engineering

## LATE interpretations of linear IV

- It holds in special cases, not general cases
- Recent textbooks discuss binary/binary case but nothing else
- Yet widely invoked in empirical literature (Blandhol et al 2022)
- *Mostly Harmless* wishcasting (Angrist & Pischke 2009, pg. 173):

  *The econometric tool remains 2SLS and the interpretation remains fundamentally similar to the basic LATE result, with a few bells and whistles ... These results provide a simple casual [sic] interpretation for 2SLS in most empirically relevant settings.*

## LATE interpretations of linear IV

- It holds in special cases, not general cases
- Recent textbooks discuss binary/binary case but nothing else
- Yet widely invoked in empirical literature (Blandhol et al 2022)
- *Mostly Harmless* wishcasting (Angrist & Pischke 2009, pg. 173):

  *The econometric tool remains 2SLS and the interpretation remains fundamentally similar to the basic LATE result, with a few bells and whistles ... These results provide a simple casual [sic] interpretation for 2SLS in most empirically relevant settings.*

## Does this matter "in practice?"

- Interesting question: reverse engineering is a purely theoretical exercise

→ Same number, different interpretation

- So the theory *is* the practice

# Outline

**Forward engineering: choose target parameter, *then* the estimator**

**Forward engineering: choose target parameter, *then* the estimator**

**Roughly five approaches**

**Forward engineering: choose target parameter, *then* the estimator**

## Roughly five approaches

1. Assume constant treatment effects (always an option!)

**Forward engineering: choose target parameter, *then* the estimator**

### Roughly five approaches

1. Assume constant treatment effects (always an option!)
2. Estimate LATEs directly

$\rightarrow$ Not often done, but there are several proposed/promising methods

**Forward engineering: choose target parameter, *then* the estimator**

### Roughly five approaches

1. Assume constant treatment effects (always an option!)
2. Estimate LATEs directly

→ Not often done, but there are several proposed/promising methods

3. Estimate selection model jointly with outcome (control function)

→ Follows a long line of literature dating to Gronau-Heckman (1974)

**Forward engineering: choose target parameter, *then* the estimator**

## Roughly five approaches

1. Assume constant treatment effects (always an option!)
2. Estimate LATEs directly

→ Not often done, but there are several proposed/promising methods

3. Estimate selection model jointly with outcome (control function)

→ Follows a long line of literature dating to Gronau-Heckman (1974)

4. Bounds that do not use a selection model

→ e.g. Manski (1994), but less often used in practice

**Forward engineering: choose target parameter, *then* the estimator**

### Roughly five approaches

1. Assume constant treatment effects (always an option!)
2. Estimate LATEs directly

$\rightarrow$ Not often done, but there are several proposed/promising methods

3. Estimate selection model jointly with outcome (control function)

$\rightarrow$ Follows a long line of literature dating to Gronau-Heckman (1974)

4. Bounds that do not use a selection model

$\rightarrow$ e.g. Manski (1994), but less often used in practice

5. Rank invariance (e.g. Chernozhukov & Hansen 2005)

$\rightarrow$ Generally viewed as too strong in practice

**Forward engineering: choose target parameter, *then* the estimator**

### Roughly five approaches

2. Estimate LATEs directly

$\rightarrow$ Not often done, but there are several proposed/promising methods

3. Estimate selection model jointly with outcome (control function)

$\rightarrow$ Follows a long line of literature dating to Gronau-Heckman (1974)

**Today I'll talk briefly about 2 and 3**

## The problem with forward engineering

- Using a parametric estimator but wanting nonparametric robustness

  (And not acknowledging implicit parametric assumptions!)
- Obvious solution: change the estimator

## The problem with forward engineering

- Using a parametric estimator but wanting nonparametric robustness

  (And not acknowledging implicit parametric assumptions!)
- Obvious solution: change the estimator

## Double/debiased machine learning (Chernozhukov et al 2018)

- One approach: use DDML to ensure rich covariates

$\rightarrow$ Estimand will be weakly causal, but still a weighted average of LATEs

## The problem with forward engineering

- Using a parametric estimator but wanting nonparametric robustness

  (And not acknowledging implicit parametric assumptions!)
- Obvious solution: change the estimator

## Double/debiased machine learning (Chernozhukov et al 2018)

- One approach: use DDML to ensure rich covariates

$\rightarrow$ Estimand will be weakly causal, but still a weighted average of LATEs

- Another approach: use DDML to estimate a more natural object:

$$\underbrace{\mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) > D_i(0)]}_{\text{unconditional LATE}} = \frac{\mathbb{E}\left[\mathbb{E}[Y_i | Z_i = 1, X_i] - \mathbb{E}[Y_i | Z_i = 0, X_i]\right]}{\mathbb{E}\left[\mathbb{E}[D_i | Z_i = 1, X_i] - \mathbb{E}[D_i | Z_i = 0, X_i]\right]}$$

- Requires estimating three functions of $X_i$ (rf/fs, plus $\mathbb{E}[Z_i | X_i]$)

## The problem with forward engineering

- Using a parametric estimator but wanting nonparametric robustness

  (And not acknowledging implicit parametric assumptions!)
- Obvious solution: change the estimator

## Double/debiased machine learning (Chernozhukov et al 2018)

- One approach: use DDML to ensure rich covariates

$\rightarrow$ Estimand will be weakly causal, but still a weighted average of LATEs

- Another approach: use DDML to estimate a more natural object:

$$\underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)]}_{\text{unconditional LATE}} = \frac{\mathbb{E}\left[\mathbb{E}[Y_i|Z_i = 1, X_i] - \mathbb{E}[Y_i|Z_i = 0, X_i]\right]}{\mathbb{E}\left[\mathbb{E}[D_i|Z_i = 1, X_i] - \mathbb{E}[D_i|Z_i = 0, X_i]\right]}$$

- Requires estimating three functions of $X_i$ (rf/fs, plus $\mathbb{E}[Z_i|X_i]$)
- `ddml` for Stata (Ahrens et al 2023) `DoubleML` for `R` (Bach et al 2021)

### Idea

- Model $Y_i(d)$ and $D_i(z)$ jointly (given $X_i$) — "control function"
- No magic here, but makes target parameters/parameterizations explicit

## Idea

- Model $Y_i(d)$ and $D_i(z)$ jointly (given $X_i$) — "control function"
- No magic here, but makes target parameters/parameterizations explicit

## Notation

- Easier to formalize with latent variable notation for selection:

$$D_i = \mathbb{1}[V_i \leq g(Z_i, X_i)] \quad \text{normalized to} \quad D_i = \mathbb{1}[U_i \leq p(Z_i, X_i)]$$

$$\text{where} \quad \underbrace{U_i \sim \text{Unif}[0,1]}_{\text{latent resistance to treatment}} \quad \text{and} \quad \underbrace{p(z,x) \equiv \mathbb{P}[D_i = 1 | Z_i = z, X_i = x]}_{\text{the propensity score}}$$

- Recall: *equivalent* to the monotonicity condition (Vytlacil, 2002)

### Idea

- Model $Y_i(d)$ and $D_i(z)$ jointly (given $X_i$) — "control function"
- No magic here, but makes target parameters/parameterizations explicit

### Notation

- Easier to formalize with latent variable notation for selection:

$$D_i = \mathbb{1}[V_i \leq g(Z_i, X_i)] \quad \text{normalized to} \quad D_i = \mathbb{1}[U_i \leq p(Z_i, X_i)]$$

$$\text{where} \quad \underbrace{U_i \sim \text{Unif}[0,1]}_{\text{latent resistance to treatment}} \quad \text{and} \quad \underbrace{p(z,x) \equiv \mathbb{P}[D_i = 1 | Z_i = z, X_i = x]}_{\text{the propensity score}}$$

- Recall: *equivalent* to the monotonicity condition (Vytlacil, 2002)

### Marginal treatment response

- $m(d|u,x) \equiv \mathbb{E}[Y_i(d)|U_i = u, X_i = x]$ — contains all (mean) information
- Organizes systematic unobservable variation in potential outcomes

### Parameterize the MTR

- Assume $m(d|u,x) = \sum_{\ell=1}^{L} \theta_\ell b_\ell(d|u,x)$ for some known functions $b$

  e.g. $m(0|u,x) = \theta_1 + \theta_2 u + \theta_3 u^2 + \theta_4' x, \quad m(1|u,x) = \theta_5 + \theta_6 u + \theta_7 u^2 + \theta_8' x$
- Identification determined by flexibility relative to instrument
- Allows for extrapolation via $u$ *if* desired

### Parameterize the MTR

- Assume $m(d|u,x) = \sum_{\ell=1}^{L} \theta_\ell b_\ell(d|u,x)$ for some known functions $b$

  e.g. $m(0|u,x) = \theta_1 + \theta_2 u + \theta_3 u^2 + \theta_4' x, \quad m(1|u,x) = \theta_5 + \theta_6 u + \theta_7 u^2 + \theta_8' x$
- Identification determined by flexibility relative to instrument
- Allows for extrapolation via $u$ *if* desired

### Linear regression with observed outcomes

$\Rightarrow$ $\mathbb{E}[Y_i|D_i, X_i, Z_i] = \sum_{\ell=1}^{L} \theta_\ell \bar{b}_\ell(D_i, X_i, Z_i)$ for identified $\bar{b}_\ell(d,x,z)$

- Simply regress $Y_i$ onto $\bar{b}(D_i, X_i, Z_i) \rightarrow$ estimate of $\theta$, and thus $m$
- Integrate appropriately to estimate your favorite target parameter
- Bootstrap for inference (generated regressor)

## Parameterize the MTR

- Assume $m(d|u,x) = \sum_{\ell=1}^{L} \theta_\ell b_\ell(d|u,x)$ for some known functions $b$
  e.g. $m(0|u,x) = \theta_1 + \theta_2 u + \theta_3 u^2 + \theta_4' x, \quad m(1|u,x) = \theta_5 + \theta_6 u + \theta_7 u^2 + \theta_8' x$
- Identification determined by flexibility relative to instrument
- Allows for extrapolation via $u$ *if* desired

## Linear regression with observed outcomes

$\Rightarrow$ $\mathbb{E}[Y_i|D_i,X_i,Z_i] = \sum_{\ell=1}^{L} \theta_\ell \bar{b}_\ell(D_i,X_i,Z_i)$ for identified $\bar{b}_\ell(d,x,z)$

- Simply regress $Y_i$ onto $\bar{b}(D_i,X_i,Z_i)$ $\rightarrow$ estimate of $\theta$, and thus $m$
- Integrate appropriately to estimate your favorite target parameter
- Bootstrap for inference (generated regressor)

## Do the integration for me!

`mtefe` for Stata (Andresen, 2018), `ivmte` for `R` (Shea & Torgovitsky, 2023)

**Marginal treatment *effect* (Heckman & Vytlacil 1999, 2005)**

- Looked at $m(1|u,x) - m(0|u,x)$ directly (subtle but key difference)
- Traditionally focused on continuous instruments, kernel estimation
- Discrete instruments: Brinch et al (2017), Mogstad et al (2018)

→ The latter paper also considers partial identification (bounds) — `ivmte`

- Linear basis (semiparametric/nonparametric) approaches easier to apply

### Marginal treatment *effect* (Heckman & Vytlacil 1999, 2005)

- Looked at $m(1|u,x) - m(0|u,x)$ directly (subtle but key difference)
- Traditionally focused on continuous instruments, kernel estimation
- Discrete instruments: Brinch et al (2017), Mogstad et al (2018)

$\rightarrow$ The latter paper also considers partial identification (bounds) — `ivmte`

- Linear basis (semiparametric/nonparametric) approaches easier to apply

### Empirical applications of MTE methods are numerous and growing

Arnold, Dobbie, Yang (2018), Kline, Walters (2016), Cornelissen, Dustmann, Raute, Schonberg (2018), Autor, Kostol, Mogstad, Setzler (2019), Ito, Ida, Tanaka (2023), Agan, Doleac, Harvey (2023), ...

## Marginal treatment *effect* (Heckman & Vytlacil 1999, 2005)

- Looked at $m(1|u,x) - m(0|u,x)$ directly (subtle but key difference)
- Traditionally focused on continuous instruments, kernel estimation
- Discrete instruments: Brinch et al (2017), Mogstad et al (2018)

→ The latter paper also considers partial identification (bounds) — `ivmte`

- Linear basis (semiparametric/nonparametric) approaches easier to apply

## Empirical applications of MTE methods are numerous and growing

Arnold, Dobbie, Yang (2018), Kline, Walters (2016), Cornelissen, Dustmann, Raute, Schonberg (2018), Autor, Kostol, Mogstad, Setzler (2019), Ito, Ida, Tanaka (2023), Agan, Doleac, Harvey (2023), ...

## Frontiers of the methodology

- Multivalued $D_i$ (e.g. Rose & Shem-Tov 2021, Norris et al 2023)
- Binary $D_i$ when monotonicity fails (e.g. Mogstad et al 2021)

### Instrumental variables with heterogeneous treatment effects

- Complicated, but we think it's worth it — many seem to agree
- Interesting that there is not so much disagreement on the *model*
- However there is disagreement on the right way to use it

**Instrumental variables with heterogeneous treatment effects**

- Complicated, but we think it's worth it — many seem to agree
- Interesting that there is not so much disagreement on the *model*
- However there is disagreement on the right way to use it

**Reverse engineering (e.g. *Mostly Harmless*)**

- Seductively easy (same estimator!)
- But doesn't actually have firm theoretical foundations
- Even when it does, the estimand is hard to interpret/transfer

### Instrumental variables with heterogeneous treatment effects

- Complicated, but we think it's worth it — many seem to agree
- Interesting that there is not so much disagreement on the *model*
- However there is disagreement on the right way to use it

### Reverse engineering (e.g. *Mostly Harmless*)

- Seductively easy (same estimator!)
- But doesn't actually have firm theoretical foundations
- Even when it does, the estimand is hard to interpret/transfer

### Forward engineering

- Selection correction/control function is most popular
- MTE for binary treatment is tested — other cases more experimental
- Or estimate LATEs directly — several methods, not well tested (yet)

# Thank you

Additional comments welcome, please email:
torgovitsky@uchicago.edu