Marginal Treatment Effects (MTE)

Mattis Gilbert

Labour Reading Group
Sciences Po, Spring 2025

Presentation based on Mogstad and Torgovitsky, 2024; Andresen, 2018 and Di Tradaglia.

Table of contents

UHTE with a selection model

- Selection model: model with built-in **TE heterogeneity** and **selection into treatment**.
- At the cost of **stronger assumptions** than standard IVs can:
  1. Estimate the **full distribution** of Treatment Effects (MTE)
  2. Back out parameters of interest. (ATE, ATT, ATUT, PRTE)
- Marginal Treatment Effect (MTE) capture heterogeneity in the TE along the unobserved dimension called **resistance to treatment**.
- Selection on gains: low vs high resistance to treatment individuals might have different gains from treatment!

The Generalized Roy Model (GRM) - setup

Model

$$Y_0 = \mu_0(X) + U_0$$
$$Y_1 = \mu_1(X) + U_1$$
$$Y = (1-D)Y_0 + DY_1$$

- Y_1 and Y_0 are **potential outcomes**
  - OHTE: functions of observables X
  - UHTE: functions of unobservables (U_0, U_1)
- Treatment effects $(Y_1 - Y_0)$ are **heterogeneous**, $ATE = \mu_1 - \mu_0$.
- $U_0 \equiv Y_0 - \mathbb{E}(Y_0|X); \quad U_1 \equiv Y_1 - \mathbb{E}(Y_1|X)$ so both are mean zero.

The Generalized Roy Model (GRM) - setup

Model

$$Y_0 = \mu_0(X) + U_0$$
$$Y_1 = \mu_1(X) + U_1$$
$$Y = (1 - D)Y_0 + DY_1$$

Assumptions

1. $D = \mathbb{1}\{V \leq \nu(X, Z)\}$
2. $Z \perp\!\!\!\perp (Y_0, Y_1, V)|X$
3. Distribution of $V|X = x$ is continuous.

- Y_1 and Y_0 are **potential outcomes**
  - OHTE: functions of observables X
  - UHTE: functions of unobservables (U_0, U_1)
- Treatment effects $(Y_1 - Y_0)$ are **heterogeneous**, $ATE = \mu_1 - \mu_0$.
- $U_0 \equiv Y_0 - \mathbb{E}(Y_0|X)$; $\quad U_1 \equiv Y_1 - \mathbb{E}(Y_1|X)$ so both are mean zero.
- Selection into treatment D depends on:
  1. Instrument / encouragement Z
  2. Heterogeneous cost / *resistance to treatment* V (RV)
- Z may not be binary; unknown function $\nu(\cdot)$

Potential outcomes & Selection models

- Generalized Roy Model **implies and is implied by** the standard assumptions in Angrist and Imbens, 1995 necessary to interpret an IV as a LATE.
- Vytlacil, 2002: Standard IV Ass. of relevance, exclusion, and monotonicity $\Leftrightarrow$ representation of a choice equation as in GRM.
- Differences are only **notational**!

Target parameters in GRM

$$\text{ATE}(x) \equiv \mathbb{E}\left[Y_1 - Y_0 \mid X = x\right] = \mu_1(x) - \mu_0(x)$$
$$\text{TOT}(x) \equiv \mathbb{E}\left[Y_1 - Y_0 \mid X = x, D = 1\right] = \mu_1(x) - \mu_0(x) + \mathbb{E}\left[U_1 - U_0 \mid X = x, D = 1\right]$$
$$\text{TUT}(x) \equiv \mathbb{E}\left[Y_1 - Y_0 \mid X = x, D = 0\right] = \mu_1(x) - \mu_0(x) + \mathbb{E}\left[U_1 - U_0 \mid X = x, D = 0\right]$$

- Same definitions as before, but now we are conditioning on X.
- Average over the distribution of X to obtain unconditional versions.
- Selection on gains appers in TOT and TUT

Normalization: Transform V to Uniform(0,1)

- $\nu(.)$ and V are unknown to econometrician.
- Yet Ass.2 (conditional full exogeneity) $\Rightarrow$ some features are identified through **propensity score**!

Normalization: Transform V to Uniform(0,1)

- $\nu(.)$ and V are unknown to econometrician.
- Yet Ass.2 (conditional full exogeneity) $\Rightarrow$ some features are identified through **propensity score**!
- Let $F_{V|X}$ be the conditional CDF of $V|X=x$ (continuous)

$$\underbrace{\pi(z,x) \equiv \mathbb{P}[D=1|Z=z,X=x]}_{\text{the (treatment) propensity score}} = \overbrace{\mathbb{P}[V \leq \nu(z,x)|X=x]}^{\text{by full exogeneity}} \equiv F_{V|X}(\nu(z,x)|x),$$

Normalization: Transform V to Uniform(0,1)

- $\nu(.)$ and V are unknown to econometrician.
- Yet Ass.2 (conditional full exogeneity) $\Rightarrow$ some features are identified through **propensity score**!
- Let $F_{V|X}$ be the conditional CDF of $V|X=x$ (continuous)

$$\underbrace{\pi(z,x) \equiv \mathbb{P}[D=1|Z=z,X=x]}_{\text{the (treatment) propensity score}} = \overbrace{\mathbb{P}[V \leq \nu(z,x)|X=x]}^{\text{by full exogeneity}} \equiv F_{V|X}(\nu(z,x)|x),$$

Then:

$$\begin{aligned} & D = \mathbb{1}\{V \leq \nu(X,Z)\} \\ \Leftrightarrow\ & D = \mathbb{1}\{\underbrace{F_{V|X}(V|X)}_{\equiv U_D} \leq \underbrace{F_{V|X}(\nu(Z,X)|X)}_{=\pi(Z,X)}\} \equiv \mathbb{1}\{U_D \leq \pi(Z,X)\}. \end{aligned}$$

- $U_D \sim Uniform(0,1)$ by Probability Integral Transform.
- $\mathbb{P}(\text{Uniform}(0,1) < c) = c$.
- **Result**: U_D is a known distribution, $\pi(X,Z)$ is *identified*.

Generalized Roy model

Model

$$Y_0 = \mu_0(X) + U_0$$
$$Y_1 = \mu_1(X) + U_1$$
$$Y = (1-D)Y_0 + DY_1$$
$$\pi(X,Z) = P(D=1|X,Z)$$

Assumptions

1. $D = \mathbb{1}\{U_D \leq \pi(X,Z)\}$
2. $Z \perp\!\!\!\perp (Y_0, Y_1, U_D)|X$
3. Distribution of $U_D|(X=x, Z=z) \sim Uniform(0,1)$

- U_D is interpretable as a **quantile** of "resistance to treatment" **conditional on** X
- It is only comparable across individuals with **same** observables (a.k.a within X)
- Indiv with lower U_D are more likely to take treatment (regardless of Z).
- **Result**: The models are observationally equivalent!
  - Do not need to know V nor $\nu(.)$
  - Only need propensity score which is *identified*!
  - A propensity score of 0.9 $\Rightarrow$ Every individual with resistance to treatment below 90^{th} percentile is treated.

Power of this normalization

This transformation is very important for identification:

$$\pi(X,Z) = P(D=1|X,Z)$$
$$D = \mathbb{1}\{U_D \leq \pi(X,Z)\}$$

- It directly links **propensity score** ($\pi(X,Z)$) to **quantiles of resistance to treatment** (U_D)
- When observing $\pi(X,Z) = .30$, it means that the 30% with lowest U_D took up treatment a.k.a $U_D \leq .30$.
- Therefore if the instrument $Z \in \{0,1\}$ shifts propensity from:

$$\pi(X,Z=0) = 0.3 \rightarrow \pi(X,Z=1) = 0.6$$

then you know that the **compliers** are the ones with

$$U_D \in [\pi(X,Z=0) = 0.3, \pi(X,Z=1) = 0.6]$$

Marginal Treatment Effects (MTEs)

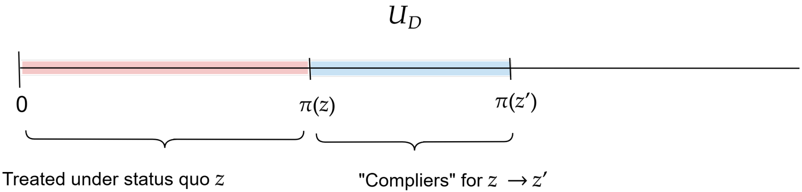
$$MTE(u,x) = \mathbb{E}(Y_1 - Y_0 | U_D = u, X = x) \quad (1)$$

$$= \underbrace{\mu_1(x) - \mu_0(x)}_{\text{Observed heterogeneity TE}} + \underbrace{\mathbb{E}[U_1 - U_0 | X = x, U_D = u]}_{\text{Unobserved heterogeneity TE}}. \quad (2)$$

- $MTE(u,x)$ is the **average causal effect** of D on Y **for individuals with selection unobservable** $U_D = u$ **and observed characteristics** $X = x$.
- MTE declining in u: indiv most likely to take treatment receive greater gains from it.
- No unobsered heterogeneity: MTE is constant (flat) along u.

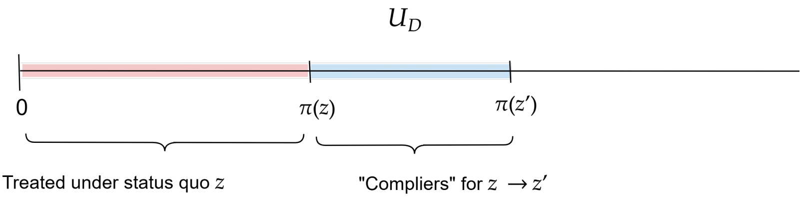
The MTE is a useful definition because it uses the selection model to **partition the population** based on **all unobservable and observable** determinants of their treatment choice ***except for the instrument, which is the source of exogenous variation***

Marginal Treatment Effects and LATE



- MTEs are closely related to LATE.
- LATE is the average effect of treatment for people who are shifted into treatment when the instrument is exogenously shifted from z to z'.
- In GRM these people (the "compliers") have U_D in the interval $[\pi(z), \pi(z')]$.

Marginal Treatment Effects and LATE



- MTEs are closely related to LATE.
- LATE is the average effect of treatment for people who are shifted into treatment when the instrument is exogenously shifted from z to z'.
- In GRM these people (the "compliers") have U_D in the interval $[\pi(z), \pi(z')]$.
- Note how, when $z - z'$ is *infinitesimally* small, so that $\pi(z') = \pi(z)$, the LATE converges to the MTE.
- MTE is thus a **limit form of LATE**.

From MTE Function to Target Parameters

Target Parameters

- ATE, TOT, TUT, LATE,PRTE, etc.

General Approach

- Any of the above (and more!) can be computed as a weighted average of the MTE.

Example: ATE from MTE

$$ATE(x) = \mathbb{E}[Y_1 - Y_0|X = x] = \mathbb{E}_{V|X=x}[MTE(X, U_D)] = \int_0^1 MTE(x, u) \times 1du$$

- Follows because $U_D|X = x \sim \text{Uniform}(0, 1)$.
- ATT: higher weights to **lower** resistance to treatment
- $ATUT$: higher weight to **higher** resistance to treatment
- Other weighting functions.

Table of contents

Ito, Ida, and Tanaka, 2023 - Electricity dynamic pricing (1/2)

Ito, Ida, and Tanaka, 2023 study dynamic electricity pricing on consumption.

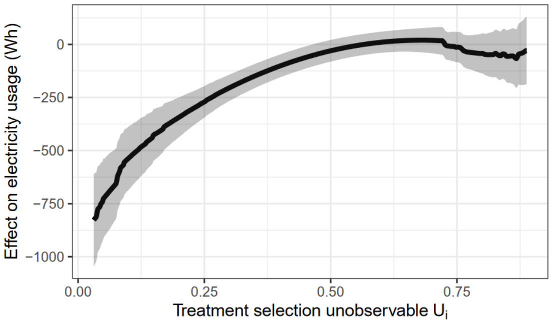
- Dynamic pricing: higher peak-hour, lower off-peak rates.
- Treatment (D): household adopts dynamic pricing.
- Instrument (Z): $60 incentive to adopt (randomly assigned).
- Outcome (Y): electricity usage.

$\Rightarrow$ Fits **LATE** setting: $D, Z \in \{0, 1\}$; Z randomly assigned (full exogeneity); satisfies relevance and monotonicity.

- LATE estimates effect of the $60 incentive.
- But what about **alternative policies**?

Ito, Ida, and Tanaka, 2023 - Electricity dynamic pricing (2/2)

***Figure 5:** Marginal treatment effect estimates from Ito et al. (2023)*



***Notes:** Authors' reproduction of Figure 10, Panel A of Ito et al. (2023). We thank Koichiro Ito for providing the necessary data. The point estimate is the estimated MTE evaluated at the sample average of the covariates. The shaded region indicates 95% bootstrapped confidence intervals.*

- Unobserved Heterogeneous Treatment Effects (UHTE):
  - "Ease of adjustment" is unobserved.
  - MTE captures UHTE via u.
- Selection on gains: consumers with higher expected savings are more likely to adopt.

Policy Implications

- Incentive size matters:
  - Small incentives attract high-impact adopters.
  - Large incentives attract more adopters with smaller effects.
- MTE relies on stronger assumptions $\Rightarrow$ Alternative policy estimates are **less credible** than LATE.
- Trade-off: LATE alone is limited, MTE provides **deeper insight**.

Cornelissen et al., 2018 - Who Benefits from Universal Child Care?

Background

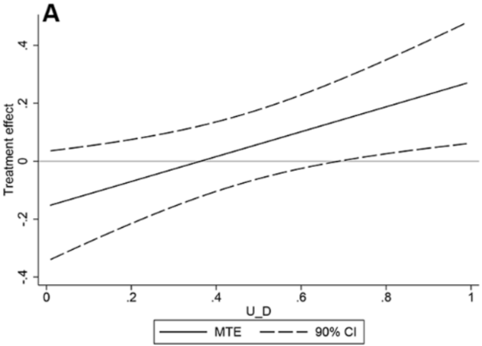
- Major policy question: causal effect of early childhood interventions, including state-provided day care.
- Some studies of highly-targeted programs (e.g. Head Start / Perry Preschool) find sizable positive effects.
- Evidence for universal provision is mixed: some find sizable negative effects (Quebec study).
- How to rationalize these conflicting findings?
- Maybe targeted programs enroll children most likely to benefit, i.e. those with an adverse home environment.

Cornelissen et al., 2018 - Who Benefits from Universal Child Care?

This Study

- Study provision of universal preschool/childcare in Germany using MTE approach.
- Use a staggered roll-out of 1990s policy reform that affected the number of slots for publicly-provided childcare in different places.
- Treatment ($D \in \{0,1\}$) is early attendance, defined as attending for at least three years.
  - Also estimate MTE for ordered selection model: $\{1,2,3\}$ years.
- Instrument (Z) is child care coverage rate in the municipality.
- Outcome (Y) is a universal school readiness exam administered at age 6.

Cornelissen et al., 2018 - Who Benefits from Universal Child Care?



- Reverse selection on gains: Minorities benefit most but enroll less.
- Similar selection on unobservables: "High resistance" children benefit most.
- Strong effect: TUT > ATE > 0 > TOT.

Ordered MTE

Table of contents

Identification

Separability (1/3)

- In principle, possible to identify MTEs with no further assumptions.
- However, would require a Z allowing $\pi(X, Z)$ to vary over the **full range** $(0, 1)$ **for any value** of X !

[0]Do not mistake unobserved potential treatment effects (U_0, U_1) with quantiles of resistance to treatment U_D

Identification

Separability (1/3)

- In principle, possible to identify MTEs with no further assumptions.
- However, would require a Z allowing $\pi(X, Z)$ to vary over the **full range** $(0, 1)$ **for any value** of X !

Assumption 4: Separability $\mathbb{E}(U_j|V, X) = \mathbb{E}(U_j|V)$ for $j \in \{0, 1\}$

[0] Do not mistake unobserved potential treatment effects (U_0, U_1) with quantiles of resistance to treatment U_D

Identification

Separability (1/3)

- In principle, possible to identify MTEs with no further assumptions.
- However, would require a Z allowing $\pi(X, Z)$ to vary over the **full range** $(0, 1)$ **for any value** of X !

Assumption 4: Separability $\mathbb{E}(U_j|V, X) = \mathbb{E}(U_j|V)$ for $j \in \{0, 1\}$

Result:

- MTEs are additively separable in U_D and X:
  - X only affects **intercept** of MTE.
  - The **pattern** of UHTE does not depend on X.
  - e.g MTE has same pattern for Men and Women but different *intercept*.
- MTE is identified over the **common support**, **unconditional on X**

[0] Do not mistake unobserved potential treatment effects (U_0, U_1) with quantiles of resistance to treatment U_D

Identification

Separability (2/3)

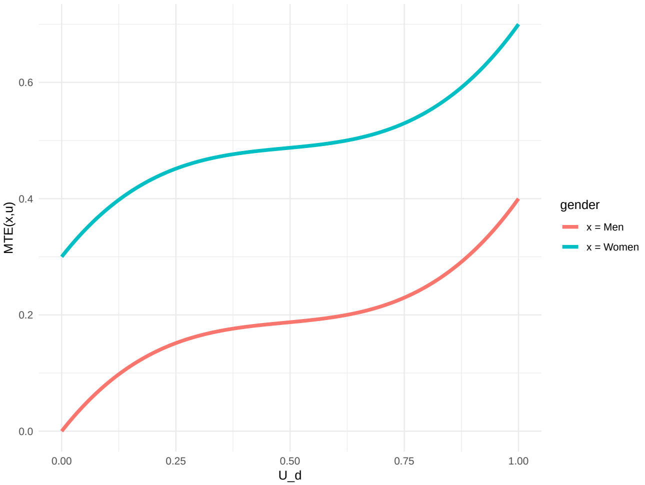


Figure: Implications of separability for MTE

Identification

Separability (3/3)

Is separability such a strong assumption?

- Not as strong as joint normality of (U_0, U_1, V) in Heckman selection model.
- Same pattern along **quantiles** of resistance to treatment, so might actually make sense.
- This assumption should be **carefully evaluated** by researcher
  - This is **application dependent**.

MTE is identified over common support of $\pi(X, Z)$

- Separability assumption $\Rightarrow$ do **not need** $\pi(X, Z)$ to have support in $(0, 1)$ for all x
  - a.k.a $supp\,(\pi(X = x, Z)) = (0, 1) \forall x$
- Instead, only need $supp\,(\pi(X, Z)) = (0, 1)$ **unconditional on** X**!**
  - MTE is identified over the **common support**!
  - Each $X = x$ will contribute to identify *some part* of the support of U_D.

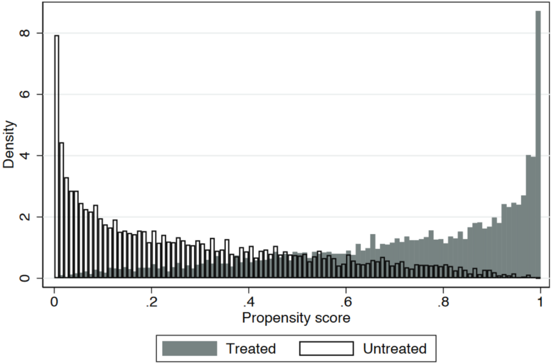


Figure: Andresen, 2018, common support plot, probit

It is common practice to assume linearity for $\mu_j(X)$:

- Restrict the way co-variates affect MTE's intercept.
- $E[Y_0|X=x] = x'\beta_0$ and $E[Y_1|X=x] = x'\beta_1$

Together, Separability + Linearity yield:

$$\begin{aligned}
\text{MTE}(u,x) &= \mu_1(x) - \mu_0(x) + E\left(U_1 - U_0 \mid X = x, U_D = u\right) \\
&= \mu_1(x) - \mu_0(x) + E\left(U_1 - U_0 \mid U_D = u\right) && \textit{(Separability)} \\
&= \underbrace{x'\left(\beta_1 - \beta_0\right)}_{\text{heterogeneity in observables}} + \underbrace{E\left(U_1 - U_0 \mid U_D = u\right)}_{k(u):\ \text{heterogeneity in unobservables}} && \textit{(Linearity)} \\
&= x'\left(\beta_1 - \beta_0\right) + k(u)
\end{aligned}$$

Note that because of Separability $k(u)$ is only a function of u, not x!

Table of contents

Estimation approaches

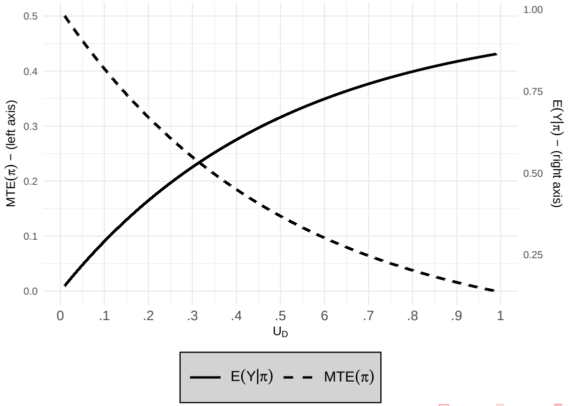
Two approaches exist to estimate MTES:

1. Local IV
   - Heckman and Vytlacil, 1999; Heckman and Vytlacil, 2001; Heckman and Vytlacil, 2005
2. Separate Approach
   - Heckman and Vytlacil, 2007; Brinch, Mogstad, and Wiswall, 2017

Table of contents

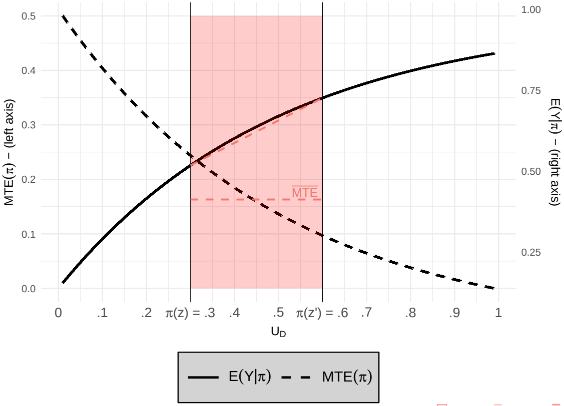
Local IV - Intuition

- For fixed X, without UHTE $\mathbb{E}(Y|\pi) = \mathbb{E}(DY_1 + (1-D)Y_0|\pi)$ is linear in π.
  - Without UHTE, only the *proportion* of treated changes, while TE is constant.
- On the other hand: with UHTE, $\mathbb{E}(Y|\pi)$ displays **non-linearities**
- LIV identifies MTE from those non-linearities!
- Hence from the **derivative** of $\mathbb{E}(Y|\pi)$ w.r.t π:
  - Constant derivative $\Rightarrow$ constant MTE $\Rightarrow$ no UHTE
  - $\uparrow / \downarrow$ derivative $\Rightarrow \uparrow / \downarrow$ MTE $\Rightarrow$ UHTE.



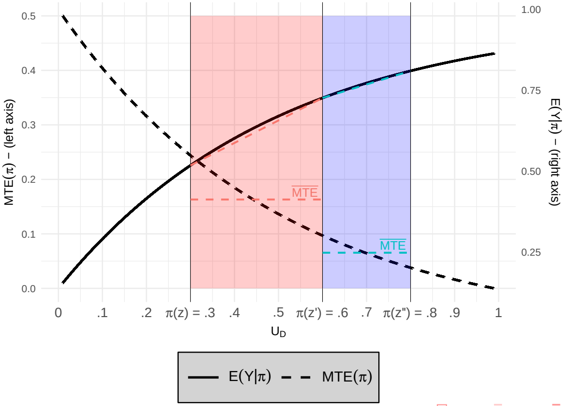
Local IV - Intuition

- For fixed X, without UHTE $\mathbb{E}(Y|\pi) = \mathbb{E}(DY_1 + (1-D)Y_0|\pi)$ is linear in π.
  - Without UHTE, only the *proportion* of treated changes, while TE is constant.
- On the other hand: with UHTE, $\mathbb{E}(Y|\pi)$ displays **non-linearities**
- LIV identifies MTE from those non-linearities!
- Hence from the **derivative** of $\mathbb{E}(Y|\pi)$ w.r.t π:
  - Constant derivative $\Rightarrow$ constant MTE $\Rightarrow$ no UHTE
  - $\uparrow / \downarrow$ derivative $\Rightarrow \uparrow / \downarrow$ MTE $\Rightarrow$ UHTE.



Local IV - Intuition

- For fixed X, without UHTE $\mathbb{E}(Y|\pi) = \mathbb{E}(DY_1 + (1-D)Y_0|\pi)$ is linear in π.
  - Without UHTE, only the *proportion* of treated changes, while TE is constant.
- On the other hand: with UHTE, $\mathbb{E}(Y|\pi)$ displays **non-linearities**
- LIV identifies MTE from those non-linearities!
- Hence from the **derivative** of $\mathbb{E}(Y|\pi)$ w.r.t π:
  - Constant derivative $\Rightarrow$ constant MTE $\Rightarrow$ no UHTE
  - $\uparrow / \downarrow$ derivative $\Rightarrow \uparrow / \downarrow$ MTE $\Rightarrow$ UHTE.

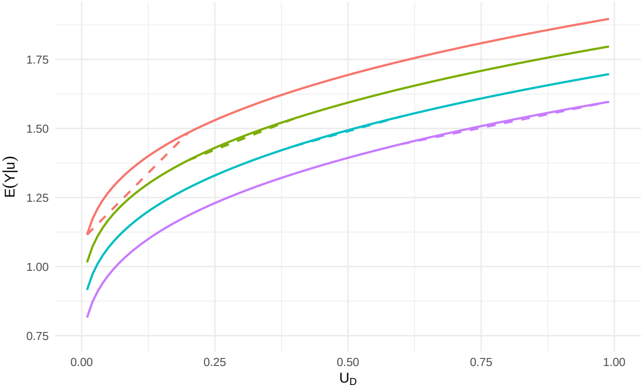


Local IV - Separability assumption

- LIV requires ≥ 3 distinct values of $\pi(Z)$ to identify linear MTE.
- Thus binary Z would not work?

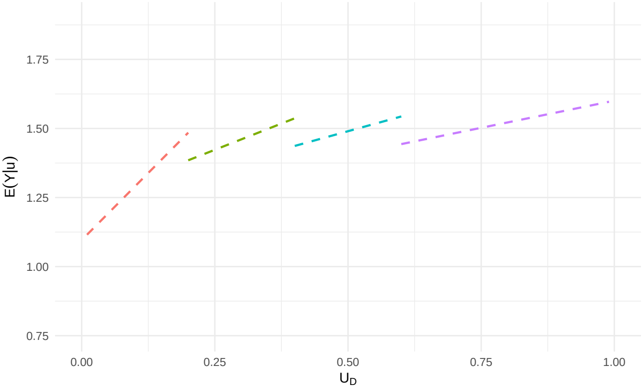
Local IV - Separability assumption

- LIV requires ≥ 3 distinct values of $\pi(Z)$ to identify linear MTE.
- Thus binary Z would not work?
- Would still work thanks to **separability assumption**!
  - Even with $Z \in \{0, 1\}$
  - Will result in different $\{\pi(z), \pi(z')\}$ across observables X!



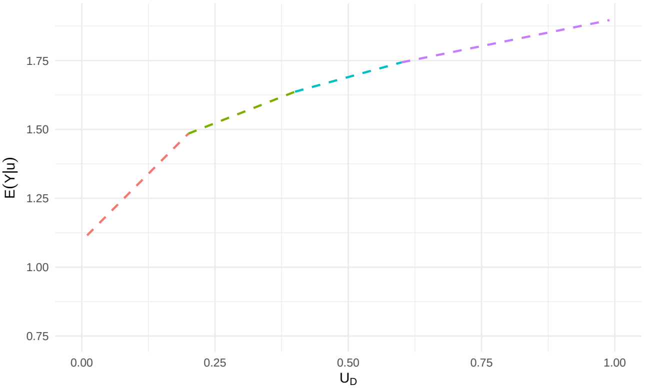
Local IV - Separability assumption

- LIV requires ≥ 3 distinct values of $\pi(Z)$ to identify linear MTE.
- Thus binary Z would not work?
- Would still work thanks to **separability assumption**!
  - Even with $Z \in \{0, 1\}$
  - Will result in different $\{\pi(z), \pi(z')\}$ across observables X!



Local IV - Separability assumption

- LIV requires ≥ 3 distinct values of $\pi(Z)$ to identify linear MTE.
- Thus binary Z would not work?
- Would still work thanks to **separability assumption**!
  - Even with $Z \in \{0, 1\}$
  - Will result in different $\{\pi(z), \pi(z')\}$ across observables X!
- Homogeneizing intercepts



Local IV - Formally

- From GRM with listed assumptions, can show that:

$$\begin{aligned}\mathbb{E}\{Y|X=x,\pi(X,Z)=p\} &= \mathbb{E}\{Y_0+D(Y_1-Y_0)|X=x,\pi(X,Z)=p\}\\ &= x\beta_0 + x(\beta_0-\beta_1)p + \underbrace{p\mathbb{E}(U_1-U_0|U_D\le p)}_{K(p)}\end{aligned}$$

- Taking the derivative of this expression with respect to p and evaluating it at u, we get the MTE:

$$\frac{\partial \mathbb{E}\{Y|X=x,\pi(X,Z)=p\}}{\partial p}\bigg|_{p=u} = x(\beta_1-\beta_0) + \frac{\partial\{p\mathbb{E}(U_1-U_0|U_D\le p)\}}{\partial p}\bigg|_{p=u}$$

$$MTE(x,u) = (\beta_1-\beta_0)x + \underbrace{\mathbb{E}(U_1-U_0|U_D=u)}_{k(u)}$$

Local IV - Estimation procedure

1. Identify selection into treatment $\pi(X, Z)$ using a **probability model**
   - e.g., probit, logit, linear probability, or semiparametric binary choice.
2. Assume a functional form for $K(p) = pE(U_1 - U_0 | U_D \leq p)$.
   - Parametric MTE (Joint normality, polynomial, polynomial w/ splines)
   - Semi-parametric (local polynomial regressions)
3. Estimate the conditional expectation of Y from eq.4

$$\mathbb{E}\{Y|X = x, \pi(X, Z) = p\} = x\beta_0 + x(\beta_0 - \beta_1)p + \underbrace{p\mathbb{E}(U_1 - U_0|U_D \leq p)}_{K(p)}$$

4. Form its derivative to obtain the MTE.

Local IV - Cornelissen et al., 2018

- Cornelissen et al., 2018 estimate parametric MTE.
- Model $K(p)$ as a **polynomial in** p

$$\mathbb{E}[Y \mid \pi(X,Z)=p, X=x] = x'\beta_0 + x'(\beta_1-\beta_0)p + \overbrace{\sum_{j=2}^{J}\alpha_j p^j}^{K(p)}$$

1. Run probit/logit of D_i on (X_i, Z_i) to estimate the propensity scores $\hat{p}_i$.
2. Estimate β_0, β_1, α from the following regression:

$$Y_i = X_i\beta_0 + X_i'(\beta_1-\beta_0)\hat{p}_i + \sum_{j=2}^{J}\alpha_j\hat{p}_i^j + \epsilon_i$$

3. Construct the estimated MTE function as follows:

$$\widehat{\text{MTE}}(p,x) = \frac{\partial}{\partial p}\left[x'\hat{\beta}_0 + x'\left(\hat{\beta}_1-\hat{\beta}_0\right)p + \sum_{j=2}^{J}\hat{\alpha}_j p^j\right]$$

4. Take weighted average of $\widehat{MTE}(p,x)$ to construct desired target parameter.

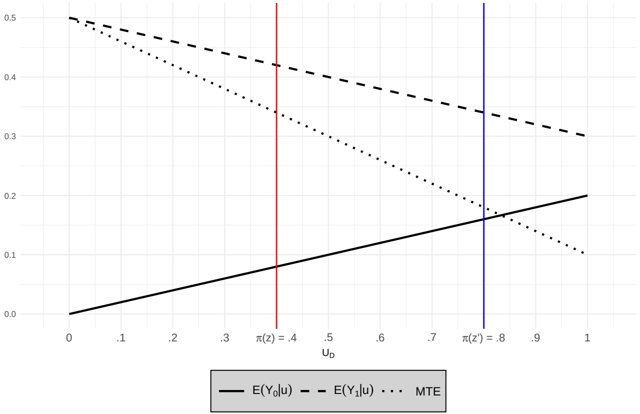
Table of contents

Separate approach - Intuition

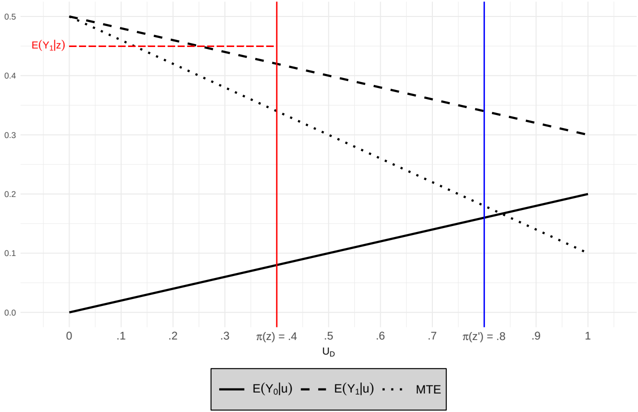
- Specify and estimate conditional expectations of Y_1 and Y_0 in treated/untreated sample **separately**

$$\mathbb{E}(Y_1|X=x, D=1) = x\beta_1 + \mathbb{E}(U_1|U_D \leq p) = x\beta_1 + K_1(p) \quad (3)$$

$$\mathbb{E}(Y_0|X=x, D=0) = x\beta_0 + \mathbb{E}(U_0|U_D > p) = x\beta_0 + K_0(p) \quad (4)$$

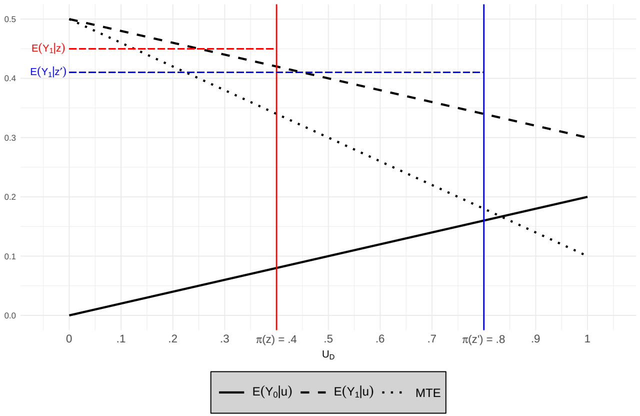


Separate approach - Intuition



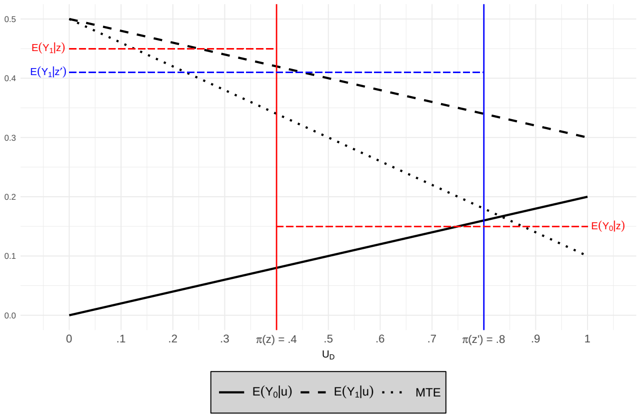
- Average outcome for **treated** under z

Separate approach - Intuition



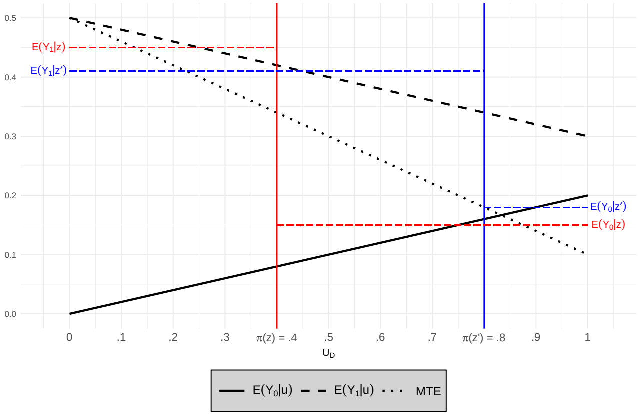
- Average outcome for **treated** under z'
- Observe **decreasing** outcome for treated
- If this trend is not **parallel** for the $Y_0|u)$, will entail **UHTE**.

Separate approach - Intuition



- Average outcome for **un-treated** under z

Separate approach - Intuition



- Average outcome for **un-treated** is **increasing** in u
- Decreasing $E(Y_1|u)$ & increasing $E(Y_1|u) \Rightarrow \downarrow$ MTE $\Rightarrow$ UHTE.
- Binary Z identifies a linear MTE model within X!

Separate approach - In practice

1. Identify selection into treatment $\pi(X, Z)$ using a **probability model**
   - e.g., probit, logit, linear probability, or semiparametric binary choice.
2. Assume a functional form for $K_j(p) = pE(U_j|U_D \leq p)$ for $j \in 0, 1$.
   - Parametric MTE (Joint normality, polynomial, polynomial w/ splines)
   - Semi-parametric (local polynomial regressions)
3. Estimate the conditional expectation of Y in the sample of treated and untreated **separately** using the regression:

$$Y_j = X\beta_j + K_j(p) + \epsilon$$

4. Form the derivatives to obtain the MTE from:

$$\begin{aligned} MTE(x,u) &= \mathbb{E}(Y_1|X=x, U_D=u) - \mathbb{E}(Y_0|X=x, U_D=u) \\ &= x(\beta_1 - \beta_0) + k_1(u) - k_0(u) \end{aligned}$$

Where $k_j(u) = \mathbb{E}(U_j|U_D = u)$

MTE in alternative Setups

- Binary D_i with violation of monotonicity
  - e.g judge FE.
  - Solution: Binarize instrument, "most" vs "least" strict. Makes less likely to violate monotonicity.
- Binary multiple instruments
  - e.g distance (close vs far) + scholarship (yes vs no).
  - Remove $(Z_{i1}, Z_{i2}) = (0, 1)$ and $(Z_{i1}, Z_{i2}) = (1, 0)$
  - Keep only $(Z_{i1}, Z_{i2}) = (0, 0)$ and $(Z_{i1}, Z_{i2}) = (1, 1)$
- Ordered treatments
  - MTE does extend to ordered treatment (e.g Bachelor vs Master)
  - But, threshold crossing not equiavelent to monotonicity anymore
  - Example : Cornelissen et al., 2018
- Continuous treatment
  - Requires extreme variations[1] in Z_i or additional assumptions.
- Unordered treatments
  - Requires extreme variations in Z_i or additional assumptions.

[1] Does not exist in practice

Table of contents

Recommendations for Practitioners (1/3)

Step 1: Assess the likely role of UHTE

- Why could treatment effects vary? Is there any reason to think that treatment effects would not vary?
- Problem occurs if UHTE is correlated with treatment choice. How plausible is it to assume that the UHTE is uncorelated with treatment choice?
- To defend non-existence of UHTE: show that there is no OHTE (precisely estimated 0s).
- If stick to constant-TE assumption: **state clearly in paper**.

Step 2: Reverse engineer with caution. The JOSH method:

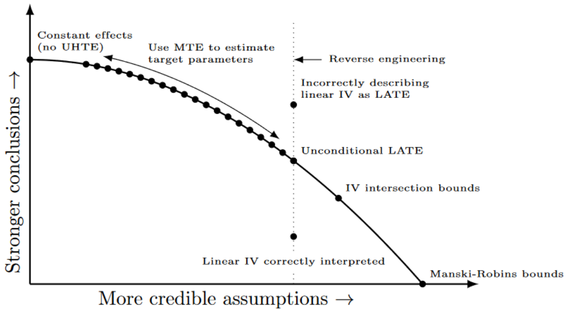
1. **Judge the setting**. Which setting is applicable? (Treatment, Instrument, Covariates)
2. **Obtain a weakly causal interpretation.** Is estimand weakly causal under full exogeneity and an appropriate monotonicity condition?
   - Primary concern = Satisfy rich covariates condition?
   - If $Z_i \perp\!\!\!\perp X_i$, automatically satisfied.
   - Otherwise: perform RESET test.
3. **Scrutinize the interpretation.** How can be the estimand interpreted? (Weights, counterfactual)
4. **Honestly communicate to the audience.** Clearly and transparently **communicate the interpretation** of the estimand and the assumptions on which the interpretation rests.

Recommendations for Practitioners (3/3)

Step 3: Forward engineer estimates of interpretable target parameters

- Choose target parameters relevant to your research question.
- What can be said about target parameters depend on **assumptions** made.
- Explore the frontier (trade-off)

Figure 7: The empirical production possibility frontier for IV methods



Notes: Two primary trade-offs involved in producing empirical research with a binary treatment.

Recommended packages

For MTE:

- STATA: mtefe (Andresen, 2018), ivmte (Shea and Torgovitsky, 2023)
- R: ivmte (Shea and Torgovitsky, 2023)

See Torgovitsky's github for more detailed list of packages.

Bibliography I

Andresen, Martin Eckhoff (2018). "Exploring marginal treatment effects: Flexible estimation using Stata". In: *The Stata Journal* 18.1, pp. 118–158.

Angrist, Joshua and Guido Imbens (1995). *Identification and estimation of local average treatment effects.*

Brinch, Christian N, Magne Mogstad, and Matthew Wiswall (2017). "Beyond LATE with a discrete instrument". In: *Journal of Political Economy* 125.4, pp. 985–1039.

Cornelissen, Thomas et al. (Dec. 2018). "Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance". In: *Journal of Political Economy* 126.6. Publisher: The University of Chicago Press, pp. 2356–2409. ISSN: 0022-3808. DOI: 10.1086/699979. URL: https://www.journals.uchicago.edu/doi/10.1086/699979 (visited on 03/24/2025).

Heckman, James J and Edward J Vytlacil (1999). "Local instrumental variables and latent variable models for identifying and bounding treatment effects". In: *Proceedings of the national Academy of Sciences* 96.8, pp. 4730–4734.

Bibliography II

- Heckman, James J and Edward J Vytlacil (2001). "Local instrumental variables". In: *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya.[Google Scholar].*
- — (2005). "Structural equations, treatment effects, and econometric policy evaluation 1". In: *Econometrica* 73.3, pp. 669–738.
- — (2007). "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments". In: *Handbook of econometrics* 6, pp. 4875–5143.
- Ito, Koichiro, Takanori Ida, and Makoto Tanaka (2023). "Selection on welfare gains: Experimental evidence from electricity plan choice". In: *American Economic Review* 113.11, pp. 2937–2973.
- Mogstad, Magne and Alexander Torgovitsky (2024). *Instrumental Variables with Unobserved Heterogeneity in Treatment Effects.* Working Paper 32927. National Bureau of Economic Research. DOI: 10.3386/w32927. URL: http://www.nber.org/papers/w32927.

Bibliography III

Shea, Joshua and Alexander Torgovitsky (2023). "ivmte: An R Package for Extrapolating Instrumental Variable Estimates Away From Compliers". In: *Observational Studies* 9.2, pp. 1–42.

Vytlacil, Edward (2002). "Independence, monotonicity, and latent index models: An equivalence result". In: *Econometrica* 70.1, pp. 331–341.

Appendix

The Generalized Roy Model (GRM) - Full exogeneity

Model

$$Y_0 = \mu_0(X) + U_0$$
$$Y_1 = \mu_1(X) + U_1$$
$$Y = (1 - D)Y_0 + DY_1$$

Assumptions

1. $D = \mathbb{1}\{V \leq \nu(X, Z)\}$
2. $Z \perp\!\!\!\perp (Y_0, Y_1, V)|X$
3. Distribution of $V|X = x$ is continuous.

- Assumption 2 implies (conditional) **full exogeneity**
- It is equivalent to $Z \perp\!\!\!\perp (Y_0, Y_1, D_0, D_1)|X$
- Z conditionally independent of Potential outcomes *and* is independent of unobservables (V) which determine potential treatment.

Back

The Generalized Roy Model (GRM) - Monotonicity

Model

$$Y_0 = \mu_0(X) + U_0$$
$$Y_1 = \mu_1(X) + U_1$$
$$Y = (1 - D)Y_0 + DY_1$$

Assumptions

1. $D = \mathbb{1}\{V \leq \nu(X, Z)\}$
2. $Z \perp\!\!\!\perp (Y_0, Y_1, V)|X$
3. Distribution of $V|X = x$ is continuous.

- Assumption 1 + 2 imply monotonicity
- Holding X fixed, we can shift $\nu(X, Z)$ by changing Z without affecting V.
- Why? Conditional on X , Z and V are independent and V doesn't enter $\nu(\cdot)$.
- For a given shift in Z, two people with the same observed characteristics X experience the same shift in $\nu(\cdot)$ *regardless of whether they have different resistance to treatment* V

Back

Policy Relevant Treatment Effects (PRTEs)

$$\text{PRTE}(x) \equiv \frac{\mathbb{E}[Y \mid X = x,\ \text{New Policy}] - \mathbb{E}[Y \mid X = x,\ \text{Old Policy}]}{\mathbb{E}[D \mid X = x,\ \text{New Policy}] - \mathbb{E}[D \mid X = x,\ \text{Old Policy}]}$$

- Compare a new policy to old one; average over X to obtain unconditional version.
- Policy $\equiv$ change in the propensity score $\pi(Z, X)$ that changes who is treated without affecting (Y_1, Y_0, V).
- PRTE is the average change in Y per person shifted into treatment.
- At some values of x, people may be shifted out of treatment
- A LATE is a PRTE, but a given LATE may not answer your policy question!

Parametric MTE

Table 1. Parametric MTE models

Function	Definition	Normal	Polynomial	Polynomial with splines
		$U_0, U_1, V \sim \mathcal{N}(0, \mathbf{\Sigma})$, $\Sigma = \left\{ \begin{matrix} \sigma_1^2 & & \\ \rho_{01} & \sigma_0^2 & \\ \rho_0 & \rho_1 & 1 \end{matrix} \right\}$	$k(u)$ or $k_j(u)$ as Lth-order polynomials with mean 0	$k(u)$ or $k_j(u)$ as Lth-order polynomials ($L \geq 2$) with Q knots for quadratic and higher-order terms at $(h_1, \ldots, h_Q)$ and mean 0
$k(u)$	$\mathbb{E}(U_1 - U_0 \mid U_D = u)$	$(\rho_1 - \rho_0)\Phi^{-1}(u)$	$\sum_{l=1}^{L} \pi_l \left(u^l - \frac{1}{l+1}\right)$	$\sum_{l=1}^{L} \pi_l \left(u^l - \frac{1}{l+1}\right) + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_k^q \left\{ (u - h_q)^l \, 1(u \geq h_q) - \frac{(1-h_q)^{l+1}}{(l+1)} \right\}$
$K(p)$	$p\mathbb{E}(U_1 - U_0 \mid U_D \leq p)$	$-(\rho_1 - \rho_0)\varphi\left\{\Phi^{-1}(p)\right\}$	$\sum_{l=1}^{L} \pi_l \frac{p(p^l - 1)}{l+1}$	$\sum_{l=1}^{L} \pi_l \frac{p(p^l+1)}{l+1} + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_k^q \left\{ \frac{1(p \geq h_q)(p-h_q)^{l+1} - (1-h_q)^{l+1} p}{l+1} \right\}$
$k_1(u)$	$\mathbb{E}(U_1 \mid U_D = u)$	$\rho_1 \Phi^{-1}(u)$	$\sum_{l=1}^{L} \pi_{1l} \left(u^l - \frac{1}{l+1}\right)$	$\sum_{l=1}^{L} \pi_{1l} \left(u^l - \frac{1}{l+1}\right) + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_{1l}^q \left\{ \mathbb{1}(u \geq h_q)(u - h_q)^l - \frac{(1-h_q)^{l+1}}{(l+1)} \right\}$
$k_0(u)$	$\mathbb{E}(U_0 \mid U_D = u)$	$\rho_0 \Phi^{-1}(u)$	$\sum_{l=1}^{L} \pi_{0l}(u^l - \frac{1}{l+1})$	$\sum_{l=1}^{L} \pi_{0l}(u^l - \frac{1}{l+1}) + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_{0l}^q \left\{ \mathbb{1}(u \geq h_q)(u - h_q)^l - \frac{(1-h_q)^{l+1}}{(l+1)} \right\}$
$K_1(p)$	$\mathbb{E}(U_1 \mid U_D \leq p)$	$-\rho_1 \frac{\varphi\left\{\Phi^{-1}(p)\right\}}{p}$	$\sum_{l=1}^{L} \pi_{1l} \frac{p^l - 1}{l+1}$	$\sum_{l=1}^{L} \pi_{1l} \frac{(p^l+1)}{l+1} + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_{1l}^q \left\{ \frac{1(p \geq h_q)(p-h_q)^{l+1} - (1-h_q)^{l+1}}{l+1} \right\}$
$K_0(p)$	$\mathbb{E}(U_0 \mid U_D > p)$	$\rho_0 \frac{\varphi\left\{\Phi^{-1}(p)\right\}}{(1-p)}$	$\sum_{l=1}^{L} \pi_{0l} \frac{p(1-p^l)}{(1-p)(l+1)}$	$\sum_{l=1}^{L} \pi_{0l} \frac{p(1-p^l)}{(1-p)(l+1)} + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_{0l}^q \left\{ \frac{(1-h_q)^{l+1} p - 1(p \geq h_q)(p-h_q)^{l+1}}{(1-p)(l+1)} \right\}$
$\text{MTE}(x, u)$	$\mathbb{E}(Y_1 - Y_0 \mid U_D = u, X = x)$		$x(\beta_1 - \beta_0) + k(u) = x(\beta_1 - \beta_0) + k_1(u) - k_0(u)$	
$Y_1(x, u)$	$\mathbb{E}(Y_1 \mid U_D = u, X = x)$		$x\beta_1 + k_1(u)$	
$Y_0(x, u)$	$\mathbb{E}(Y_0 \mid U_D = u, X = x)$		$x\beta_0 + k_0(u)$	

Note: Expressions for conditional expectations with different assumptions for the joint distribution of the error terms. $\mathbb{1}(A)$ is the indicator function for the event A. Note that $\pi_l = \pi_{1l} - \pi_{0l}$ and equivalently for the spline coefficients. Calculated as $K(p) = \int_0^p k(u)du$, $K_1(p) = 1/p \int_0^p k_1(u)du$ and $K_0(p) = 1/(1-p) \int_p^1 k_0(u)du$.

Semi-parametric MTE

Table 2. Semiparametric MTE models

Step	Local IV	Separate approach
Estimating equation	$Y = X\beta_0 + X(\beta_1 - \beta_0) + K(p) + \epsilon$	$Y_j = X\beta_j + K_j(p) + \epsilon$
Double residual regression (Robinson 1988)	local polynomial regressions of Y, X, and $X \times p$ give residuals e_Y, e_X, and $e_{X\times P}$	separate local polynomial regressions of Y and X on p in treated and untreated samples give residuals e_{Y_j} and e_{X_j}, construct $e_Y = De_{Y_1} + (1-D)e_{Y_0}$ and similar for e_X
Estimate β_0, $\beta_1 - \beta_0$ using regression	$e_Y = e_X\beta_0 + e_{X\times p}(\beta_1 - \beta_0) + \epsilon$	$e_Y = e_X\beta_0 + D(\beta_1 - \beta_0)e_X + \epsilon$
Construct residual	$\widetilde{Y} = Y - X\widehat{\beta_0} - X\left(\widehat{\beta_1 - \beta_0}\right)p$	$\widetilde{Y} = Y - X\widehat{\beta_0} - X\left(\widehat{\beta_1 - \beta_0}\right)D$
Estimate K	local polynomial regression of $\widetilde{Y}$on p, saving level $\widehat{K(p)}$ and slope $\widehat{K'(p)}$	separate local polynomial regressions of $\widetilde{Y}$ on p in treated and untreated samples, saving level $\widehat{K_j(p)}$ and slope $\widehat{K_j'(p)}$
Construct k	$\widehat{k(u)} = \widehat{K'(p)}$	$\widehat{k_1(u)} = \widehat{K_1(p)} + p\widehat{K_1'(p)}$ $\widehat{k_0(u)} = \widehat{K_0(p)} - (1-p)\widehat{K_0'(p)}$
Construct MTE	$\widehat{\text{MTE}}(x,u) = x\left(\widehat{\beta_1 - \beta_0}\right) + \widehat{k(u)}$	$\widehat{\text{MTE}}(x,u) = x\left(\widehat{\beta_1 - \beta_0}\right) + \widehat{k_1(u)} - \widehat{k_0(u)}$

Note: Steps in the estimation of semiparametric MTE models using local IVs or the separate approach. To see the relation between $k_j(u)$ and $K_j(p)$, note that $K_1(p) = \mathbb{E}(U_1|U_D \leq p) = 1/p\int_0^p \mathbb{E}(U_1|U_D = u)du \Rightarrow K_1'(p) = -(1/p)K_1(p) + (1/p)k_1(u)$, which leads to $k_1(u) = K_1(p) + pK_1'(p)$. We can find similar expressions for $k_0(u)$.

In principle, it is possible to combine the semiparametric and the polynomial approach by first estimating the β coefficients from the polynomial model and then using semiparametric methods to find K. This is the semiparametric polynomial MTE model, implemented if `polynomial()` and `semiparametric` are specified together in `mtefe`. Although computationally far less complex, there is little theory to think that the semiparametric estimate of this model should be any better than the MTE constructed from the parametric estimates.

Back

Target parameters: MTR weights

Table 5: Marginal treatment response weights for common target parameters

		MTR weights	
Target parameter	**Expression**	$\omega(1\|u,z,x)$	$\omega(0\|u,z,x)$
Average treated outcome	$\mathbb{E}[Y_i(1)]$	1	0
Average untreated outcome	$\mathbb{E}[Y_i(0)]$	0	1
Average treatment effect (ATE)	$\mathbb{E}[Y_i(1)-Y_i(0)]$	1	-1
Conditional ATE	$\mathbb{E}[Y_i(1)-Y_i(0)\|X_i \in \mathcal{X}]$	$\frac{\mathbb{1}[x \in \mathcal{X}]}{\mathbb{P}[X_i \in \mathcal{X}]}$	$-\omega(1\|u,z,x)$
Average treatment on the treated (ATT)	$\mathbb{E}[Y_i(1)-Y_i(0)\|D_i=1]$	$\frac{\mathbb{1}[u \le p(z,x)]}{\mathbb{P}[D_i=1]}$	$-\omega(1\|u,z,x)$
Average treatment on the untreated (ATU)	$\mathbb{E}[Y_i(1)-Y_i(0)\|D_i=0]$	$\frac{\mathbb{1}[u > p(z,x)]}{\mathbb{P}[D_i=0]}$	$-\omega(1\|u,z,x)$
Generalization of the LATE to $U_i \in [\underline{u},\overline{u}]$	$\mathbb{E}[Y_i(1)-Y_i(0)\|U_i \in [\underline{u},\overline{u}]]$	$\frac{\mathbb{1}[\underline{u} < u \le \overline{u}]}{\overline{u}-\underline{u}}$	$-\omega(1\|u,z,x)$
Average selection on treatment effects	$\mathbb{E}[Y_i(1)-Y_i(0)\|D_i=1] - \mathbb{E}[Y_i(1)-Y_i(0)\|D_i=0]$	$\frac{\mathbb{1}[u \le p(z,x)]}{\mathbb{P}[D_i=1]} - \frac{\mathbb{1}[u > p(z,x)]}{\mathbb{P}[D_i=0]}$	$-\omega(1\|u,z,x)$
Average selection bias	$\mathbb{E}[Y_i(0)\|D_i=1] - \mathbb{E}[Y_i(0)\|D_i=0]$	$\frac{\mathbb{1}[u \le p(z,x)]}{\mathbb{P}[D_i=1]} - \frac{\mathbb{1}[u > p(z,x)]}{\mathbb{P}[D_i=0]}$	0
Policy relevant treatment effect (PRTE)	$\frac{\mathbb{E}[Y_i^\circ]-\mathbb{E}[Y_i]}{\mathbb{E}[D_i^\circ]-\mathbb{E}[D_i]}$	$\frac{\mathbb{P}[p^\circ(X_i,Z_i^\circ) \ge u] - \mathbb{P}[p(X_i,Z_i) \ge u]}{\mathbb{E}[p^\circ(X_i,Z_i^\circ)] - \mathbb{E}[p(X_i,Z_i)]}$	$-\omega(1\|u,x)$

Notes: *The weights show how to produce the specified target parameter through the formula*

$$\textit{target parameter} = \mathbb{E}\left[\int_0^1 \text{MTR}(1|u,X_i)\omega(1|u,Z_i,X_i)\,\mathrm{d}u + \int_0^1 \text{MTR}(0|u,X_i)\omega(0|u,Z_i,X_i)\,\mathrm{d}u\right].$$

▸ Back

- When weights are symetric (a.k.a $\omega(1|u,Z_i,X_i) = -\omega(0|u,Z_i,X_i)$):

$$\text{target parameter} = \mathbb{E}\left[(\text{MTR}(1|u,X_i) - \text{MTR}(0|u,X_i))\,\omega(1|u,Z_i,X_i)\right]$$

Target parameters: MTE weights

Table 3. Unconditional treatment-effect parameters and weights

	Parameter	Event A	$\widehat{\kappa}_i$	$\widehat{\omega}(u)$
ATE	ATE	1	1	$\frac{1}{s}$
ATT	ATT	$U_D \leq p$	$\frac{p_i}{\mathbb{E}(p)}$	$\frac{P(p>u)}{s\mathbb{E}(p)}$
ATUT	ATUT	$U_D > p$	$\frac{1-p_i}{1-\mathbb{E}(p)}$	$\frac{1-P(p>u)}{s\{1-\mathbb{E}(p)\}}$
LATE	Local ATE	$P(z) < U_D \leq P(z')$	$\frac{p(z_i)-p(z_i')}{\mathbb{E}(p')-\mathbb{E}(p)}$	$\frac{P\{p(z')>u\}-P\{p(z)>u\}}{s(\overline{p'}-\overline{p})}$
2SLS	Weighted average LATE		$\frac{\{\widehat{v}_i-\mathbb{E}(\widehat{v})\}(D_i-\overline{D})}{\text{cov}(D,\widehat{v})}$	$\frac{\{\mathbb{E}(\widehat{v}\|p>u)-\mathbb{E}(\widehat{v})\}P(p>u)}{s\times\text{cov}(D,\widehat{v})}$
PRTE	Policy-relevant ATE	$p < U_D \leq p'$	$\frac{p_i'-p_i}{\mathbb{E}(p')-\mathbb{E}(p)}$	$\frac{P(p'>u)-P(p>u)}{s\{\mathbb{E}(p')-\mathbb{E}(p)\}}$
MPRTE1	Marginal PRTE 1	$\|\gamma Z - V\| < \epsilon$	$\frac{f_V(\gamma Z)}{\mathbb{E}\{f_V(\gamma Z)\}}$	$\frac{f_p(u)f_V\{F_V^{-1}(u)\}}{\mathbb{E}\{f_V(\gamma Z)\}}$
MPRTE2	Marginal PRTE 2	$\|p - U\| < \epsilon$	1	$f_p(u)$
MPRTE3	Marginal PRTE 3	$\|\frac{p}{U} - 1\| < \epsilon$	1	$\frac{uf_p(u)}{\mathbb{E}(p)}$

Note: Weights for common treatment effects. Discrete distribution of U_D with s points.

Back

Cornelissen et al., 2018 - Ordered MTE

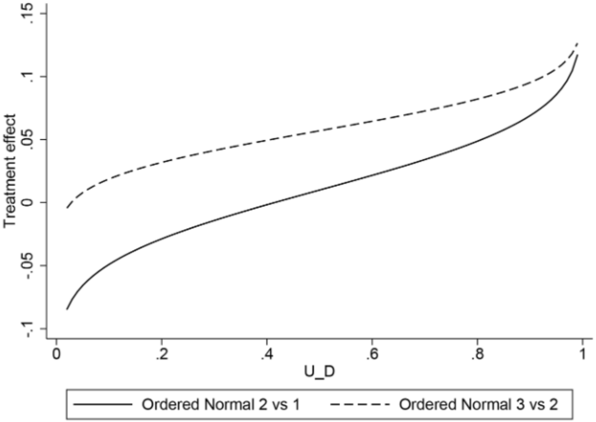
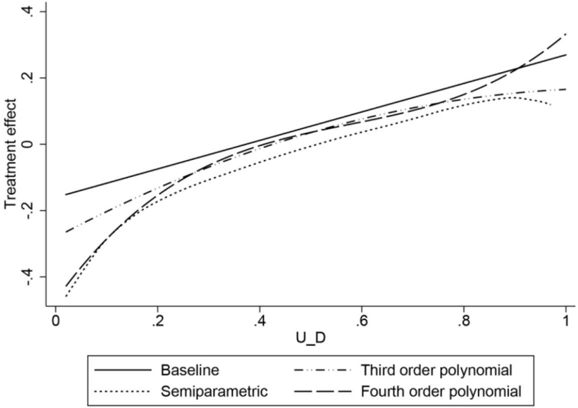


Fig. 6.—MTE curves from ordered selection model for 1, 2, and 3 years of child care. The figure displays separate MTE curves for the effects of moving from 1 year of child care to 2 years of child care (solid line) and moving from 2 years of child care to 3 years (dashed line) for the outcome of school readiness based on a normal selection model with a generalized ordered probit selection equation (see app. B for a description of that model). The curves are evaluated at mean values of the covariates. Both curves are statistically significantly upward sloping pointing toward a selection pattern of reverse selection on gains. Source: Authors' calculations based on school entry examinations, Weser-Ems, 1994–2002, as the main data source.

Cornelissen et al., 2018 et al MTE - robustness polynomials



Reverse Engineering, context

Table 3: Reverse engineering linear IV estimands

Treatment (D_i)	Instruments (Z_i)	Covariates (X_i)	Summary
Bin.	Bin.	No	The Wald and simple linear IV estimands are equal to each other and equal to the LATE under monotonicity and full exogeneity.
Bin.	Mul.	No	Each pair of instrument values defines a different complier group with an associated LATE. Different linear IV estimands produce different weighted averages of LATEs. 2SLS with a saturated instrument specification leads to non-negative weights. The weights can be negative with non-saturated specifications, but will be non-negative if the specification reproduces the monotonicity order of the instruments. The monotonicity condition can be especially unattractive if the multivalued instrument is not ordered, for example in judge designs, or when there are multiple instruments.
Ord.	Any	No	If the instrument is binary, and the treatment is a single scalar cardinal variable, then the linear IV estimand can be interpreted as the average causal response (ACR). The ACR can in turn be interpreted either as an average treatment effect among overlapping groups whose treatment choice is shifted by the instrument, or an average per-unit treatment effect across all (disjoint) complier groups. The second interpretation is a natural generalization of the LATE from the binary treatment case. If the instrument is multivalued, then these generalized LATEs get averaged according to different instrument contrasts, the same way as in the binary treatment case, and with the same caveats. Ordered treatments that are not cardinal are better analyzed through the unordered treatment case.
Uno.	Any	No	The linear IV estimand in this case has indicators for each treatment state, except for the excluded state, which is captured by a constant. If there are instruments that affect each treatment state, then the two linear IV estimands will be weakly causal if and only if each instrument affects choices only in its targeted treatment state and the excluded state is always the preferred or next best choice. Achieving this requires strong behavioral restrictions or data on next best choices. With ordered treatments that are not cardinal there are possibilities for restoring a weakly causal interpretation, but they are complicated; see main text.
Any	Any	Yes	Two assumptions are required for a linear IV estimand to be interpretable as a convex weighted average of LATEs: rich covariates and a monotonicity-correct first stage. Rich covariates is often satisfied in randomized experiments, but may not be satisfied when the instrument is not independent of covariates. The first stage will usually be monotonicity-correct under strong monotonicity, but under weak monotonicity it will only be monotonicity-correct if it includes covariates in a way that is flexible enough to account for changes in the direction of monotonicity across covariates.

Notes: *This table summarizes the discussion in Section 3.*

MLE process

Relevant only for the joint normal model, the **mlikelihood** option implements the maximum likelihood estimator described in Lokshin and Sajaia (2004). The individual log-likelihood contribution is

$$\ell_i = D_i \left[\ln\left\{F(\eta_{1i})\right\} + \ln\left\{\frac{1}{\sigma_1} f\left(\frac{U_{1i}}{\sigma_1}\right)\right\}\right]$$
$$+ (1 - D_i)\left[\ln\left\{F(-\eta_{0i})\right\} + \ln\left\{\frac{1}{\sigma_0} f\left(\frac{U_{0i}}{\sigma_{10}}\right)\right\}\right]$$
$$\text{where } \eta_{ji} = \left(\gamma Z_i + \rho_j \frac{U_{ji}}{\sigma_j}\right) \frac{1}{\sqrt{1-\rho_j^2}}$$

where f is the standard normal density. This log likelihood can be maximized to give the coefficients $\gamma, \beta_0, \beta_1, \sigma_0, \sigma_1, \rho_0$, and ρ_1. These parameter estimates can be used to construct the MTE and treatment-effect parameters as detailed in table 1.

MTE and Marginal Treatment Responses (MTR)

- MTE can be decomposed into 2 conditional expectations called MTR.

$$\text{MTR}(d|u,x) \equiv \mathbb{E}[Y(d)|U=u, X=x] \tag{5}$$

- Any target parameter that reflects a mean or a mean contrast of potential outcomes can be written as a weighted average of the MTR function.
- For isntance, the ATE

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}_X\left[\int_0^1 (\text{MTR}(1|u,X) - \text{MTR}(0|u,X))du\right] \tag{6}$$

Key to these weighting expressions is that the weights themselves are identified.

Separate approach - estimation

1. Construct $K_1(p)$ and $K_0(p)$, depending on your specification; see tables 1 and 2.
2. Estimate the conditional mean of Y from the stacked regression:

$$Y = X\beta_0 + K_D(p) + D\{X(\beta_1 - \beta_0) + K_D(p)\} + \epsilon$$

   where

$$K_D(p) = \begin{cases} K_1(p) & \text{if } D = 1 \\ K_0(p) & \text{if } D = 0 \end{cases}$$

3. From these estimates, recover the $k_j(u)$ functions.
4. Construct the estimates of the potential outcomes and the MTE as

$$\hat{Y}_j(x,u) = x\hat{\beta}_j + \widehat{k_j(u)}$$

$$\widehat{MTE(x},u) = \hat{Y}_1(x,u) - \hat{Y}_0(x,u)$$