Causal Inference
4 - Marginal Treatment Effects

Benjamin Elsner
benjamin.elsner@ucd.ie

Marginal Treatment Effects

Main Sources:

Carneiro *et al.* (2011), Cornelissen *et al.* (2016)
Mogstad *et al.* (2018), Mogstad & Torgovitsky (2018)

Marginal Treatment Effects (MTE)

MTE is a (very important and useful) **generalization of IV**

Point of departure: **LATE**

- Remember: with **heterogeneous treatment effects**...
- IV estimates the **average treatment effect for the compliers** (LATE)

The estimate depends on the **group of compliers** whose behavior is **changed by the instrument**

- different instruments give you different results
- even if all instruments are valid
- ...because the **complier population may differ between instruments**

Marginal Treatment Effects (MTE)

LATE is often not an **interesting policy parameter**

- it is defined by the group of compliers
- unless the instrument *is* the policy, LATE is not determined by the policy
- it may not coincide with the **Target Parameter**

Potentially policy-relevant parameters:

- Average Treatment Effect (ATE)
- Average Treatment Effect on the Treated (ATT)
- Average Treatment Effect on the Untreated (ATU)
- Policy-relevant Treatment Effect (PRTE)

PRTE: effect on a subpopulation that is **shifted into treatment by a policy** (more later)

Set-up

Binary treatment $D \in \{0, 1\}$

Observed vs. potential outcomes

$$Y_i = DY_{i1} + (1 - D)Y_{i0}$$

Instrument Z affects the likelihood of choosing the treatment

Potential Outcomes: The Roy Model

Assume unobserved and observed determinants are separable

$$Y_{i1} = \mu_1 + U_{i1}$$
$$Y_{i0} = \mu_0 + U_{i0}$$

With $E(U_{ij}) = 0$

$$\Delta_i = Y_{i1} - Y_{i0} = \underbrace{\mu_1 - \mu_0}_{\text{observed}} + \underbrace{U_{i1} - U_{i0}}_{\text{unobserved}}$$

Treatment Effect: Common and Individual Gains from Treatment

The **individual treatment effect** has an **important interpretation**

$$\Delta_i = Y_{i1} - Y_{i0} = \underbrace{\mu_1 - \mu_0}_{\text{observed}} + \underbrace{U_{i1} - U_{i0}}_{\text{unobserved}}$$

$\mu_1 - \mu_0$ is the **common gain from treatment** for all individuals

$U_{i1} - U_{i0}$ is the **individual gain from treatment**. This may vary across individuals.

Modelling Treatment Choice (linear case)

Latent net benefit of choosing the (binary) treatment

$$D^* = \alpha + \beta Z_i + V_i$$

Z_i	observed, **instrument(s)**
$-V_i$	unobserved **resistance to treatment**
	continuously distributed with $E(V_i = 0)$

It is possible to add |observed determinants of treatment choice X to the above equation

Treatment is chosen if the **net gain is positive**

$$D = 1 \text{ if } D^* \geq 0, D = 0 \text{ otherwise}$$

Note: D^* is unobservable

The Corresponding Regression Model

Suppose we want to **estimate** $E(\Delta_i)$ using the equation

$$Y_i = \mu_0 + \Delta_i D_i + U_{i0}.$$

We know that the **individual treatment effect** is

$$\begin{aligned} \Delta_i = Y_{i1} - Y_{i0} &= \mu_1 - \mu_0 && + U_{i1} - U_{i0} \\ &= E(\Delta_i) && + U_{i1} - U_{i0} \end{aligned}$$

Therefore, the **regression equation** becomes

$$\begin{aligned} Y_i &= \mu_0 + E(\Delta_i) D_i + U_{i0} + D_i[U_{i1} - U_{i0}] \\ &= \mu_0 + E(\Delta_i) D_i + \varepsilon_i. \end{aligned}$$

In an OLS regression, the coefficient of interest is $E(\Delta_i)$

Selection Bias

The problem here is that the treatment also determines the error term ε_i and, thus, $E(\varepsilon_i D_i) \neq 0$

More formally $E(\varepsilon_i D_i) = E(U_{i1}|D_i = 1) * Pr(D_i = 1)$

If selection into treatment is not random, the estimation of the average treatment effect $E(\Delta_i)$ suffers from **selection bias**

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \underbrace{E(\Delta_i)}_{\text{True ATE}} + \underbrace{E(U_{i1}|D_i = 1) - E(U_{i0}|D_i = 0)}_{\text{Selection Bias}}$$

Selection Bias

Selection bias has two components:

- $E(U_{i1}|D_i = 1)$ the unobservable outcome of the treated in case of treatment
- $E(U_{i0}|D_i = 0)$ the unobservable outcome of the untreated in case of no treatment

These expectations are the same under two conditions:

- Treatment was randomly assigned
- Compliance with the assignment is 100%

If **units select into treatment**, we **cannot consistently estimate the ATE** due to selection bias!

Implications of Selection Bias

In **randomized experiments**:

- None if compliance is perfect, i.e. $Pr(D_i = 1|Z_i) = 1$
- With imperfect compliance $Pr(D_i = 1|Z_i) < 1$: we still estimate a causal effect, but not the ATE.
- Is the ATE such an interesting parameter? We'll see later...

In **observational studies**:

- We have an omitted variable bias unless we find a credible identification strategy (e.g. a valid instrument)
- And even with a credible identification strategy, we may not consistently estimate the ATE

Overcoming Selection Bias: Control Functions

Heckman (1978) developed a method to **overcome selection bias**

- Idea: **estimate the extent of selection bias** and control for it

We start with constant treatment effects, $U_{i1} = U_{i0}$, $\Delta = \mu_1 - \mu_0$ and a model with covariates X_i

$$Y_i = \mu + \gamma X_i + \Delta D_i + U_i$$

$$D^* = \alpha + \beta Z_i + V_i$$

$$D_i = \begin{cases} 1 \text{ if } D_i^* \geq 0 \\ 0 \text{ if } D_i^* < 0 \end{cases}$$

Assume: $E(U_i) = E(V_i) = 0$ but $Cov(D_i, U_i) \neq 0$ (**endogenous dummy variable model**)

Switching Regression Model

Regime $D_i = 1$	$D_i^* \geq 0$	$V_i \geq -\alpha - \beta Z_i$	$Y_i = \mu + \gamma X_i + \Delta + U_i$
Regime $D_i = 0$	$D_i^* < 0$	$V_i < -\alpha - \beta Z_i$	$Y_i = \mu + \gamma X_i + U_i$

Because of **endogenous selection into treatment** ($cov(D_i, U_i) \neq 0$, the expectation of the error terms U_i *within each regime* is not zero.

This means that the **regression is misspecified** and our estimates for Δ are inconsistent

$$E(U_i | V_i \geq -\alpha - \beta Z_i) \neq E(U_i) = 0$$
$$E(U_i | V_i < -\alpha - \beta Z_i) \neq E(U_i) = 0$$

Heckman (1978) shows how these expectations can be estimated based on **distributional assumptions**

Key Ingredient: Truncated Normal Distributions

Assume that U and V are **jointly normally distributed** with

- mean zero $E(U) = E(V) = 0$
- standard deviations σ_U and σ_V
- covariance σ_{UV}
- Pearson correlation ρ_{UV}

Let $\phi(\cdot)$ be the standard normal density and $\Phi(\cdot)$ its CDF. Some useful properties of **truncated normal distributions** are:

$$E\left\{\frac{U}{\sigma_U} \mid \frac{U}{\sigma_U} > a\right\} = \frac{\phi(a)}{1-\Phi(a)}$$

$$E\left\{\frac{U}{\sigma_U} \mid \frac{U}{\sigma_U} < b\right\} = -\frac{\phi(b)}{\Phi(b)}$$

$$E\left\{\frac{U}{\sigma_U} \mid a < \frac{U}{\sigma_U} < b\right\} = \frac{\phi(a)-\phi(b)}{\Phi(b)-\Phi(a)}$$

The expressions on the right are known as the **Inverse Mills' Ratios**

Key Ingredient: Truncated Normal Distributions

Important properties of truncated joint normal distributions:

$$E\left\{\frac{U}{\sigma_U} \mid \frac{V}{\sigma_V} > a\right\} = \rho_{UV} E\left\{\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} > a\right\}$$
$$= \sigma_{UV} \frac{\phi(a)}{1 - \Phi(a)}$$

$$E\left\{\frac{U}{\sigma_U} \mid \frac{V}{\sigma_V} < a\right\} = -\rho_{UV} E\left\{\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} < a\right\}$$
$$= -\sigma_{UV} \frac{\phi(a)}{\Phi(a)}$$

$\Rightarrow$ based on parametric assumptions about the joint distribution of U and V, we can obtain estimators for selectivity.

Heckman's Two-step Procedure

Step 1: estimate the **degree of selectivity**

- Use probit to estimate α and β of the participation equation $D^* = \alpha + \beta Z_i + V_i$
- Calculate the **Inverse Mills Ratio** (IMR) for each regime

$$\lambda_1 = \frac{\phi\left(-\hat{\alpha} - \hat{\beta} Z_i\right)}{1 - \Phi\left(-\hat{\alpha} - \hat{\beta} Z_i\right)} = \frac{\phi\left(\hat{\alpha} + \hat{\beta} Z_i\right)}{\Phi\left(\hat{\alpha} + \hat{\beta} Z_i\right)}$$

$$\lambda_0 = \frac{\phi\left(-\hat{\alpha} - \hat{\beta} Z_i\right)}{\Phi\left(-\hat{\alpha} - \hat{\beta} Z_i\right)} = \frac{\phi\left(\hat{\alpha} + \hat{\beta} Z_i\right)}{1 - \Phi\left(\hat{\alpha} + \hat{\beta} Z_i\right)}$$

Heckman's Two-step Procedure

Step 2: include the IMRs into the regression equations for each regime:

$$Y_i = \mu + \gamma X_i + \Delta + \rho_1 \lambda_{1i} + \nu_i$$
$$Y_i = \mu + \gamma X_i + \rho_0 \lambda_{0i} + \nu_i$$

We can estimate both with OLS and obtain the estimate for Δ by comparing the two constants. This can also be estimated with maximum likelihood.

Because the inclusion of λ_1, λ_0 controls for selectivity as a function of unobserved resistance to treatment, it is also called a **control function**

Note: ρ_1 and ρ_0 are estimators for $\sigma_U \sigma_{UV}$ and $-\sigma_U \sigma_{UV}$, respectively

Summary: Heckman's Two-step Procedure

The Heckman (1978) method provides a first (?) **solution for eliminating selection bias**

It explicitly **models selection into treatment**

And provides a statistical methodology for estimating **unobservable selection into treatment**

We can obtain **consistent estimates for the ATE** by including a **control function** for selectivity

Summary: Heckman's Two-step Procedure

It is **important to have a valid instrument**

- The instrument must not have a direct effect on the outcome
- Finding such an instrument is challenging...

Most software packages have **in-built procedures**

In related work, Heckman (1979) applied this method to **sample selection as a specification error**

This method is less commonly used today, but it is important for understanding IV and MTE!

What Effect are We really after?

So far, the goal was to **consistently estimate the ATE**

- What is the effect of treatment for the average person...
- is this really so interesting?
- And having heard about LATE, can we actually obtain an estimate for the ATE?

Researchers need to **take a stand on the population** to which their findings apply

- Common discrepancy: the **population of interest** vs...
- the population about which our **data is informative**
- LATE is a case in point $\Rightarrow$ how interesting are compliers?

Just because an effect is **causal doesn't mean it's interesting**...

Heterogeneous Treatment Effects

In most cases, the **treatment effects** (i.e. returns to treatment) **vary** in the population

Examples:

- Free childcare: rich vs. poor families, older vs. younger people
- Job training: high- vs low-skilled workers, older vs. younger workers, etc

If units cannot be forced to be treated, we should expect those to take the **treatment for whom it is most beneficial**

Heterogeneous Treatment Effects

Bjorklund & Moffitt (1987) incorporate **heterogeneous treatment effects** into a **selection model**

$$
\begin{aligned}
Y_i &= X_i\beta + \alpha_i D_i + \varepsilon \quad \text{with} \quad D_i = 1 \text{ if } D_i^* > 0 \\
D_i^* &= \alpha_i - \phi_i \\
\alpha_i &= Z_i\delta + u_i \\
\phi_i &= W_i\eta + v_i \\
E(\varepsilon_i) &= E(u_i) = E(v_i) = 0
\end{aligned}
$$

Idea:

- Units select into treatment if the gains α_i are greater than the costs ϕ_i
- Z_i and W_i are shifters of the gains and costs (i.e. potential instruments)
- Gains and costs differ across units, $\sigma_u, \sigma_v > 0$
- It is possible that ε_i, u_i and v_i are correlated

Marginal Treatment Effects: Bjorklund & Moffitt (1987)

The **reduced form of the model** is

$$
\begin{aligned}
Y_i &= X_i\beta + Z_i\delta + \epsilon_i + u_i, \quad \text{if } D_i = 1 \\
Y_i &= X_i\beta \quad + \epsilon_i \quad , \quad \text{if } D_i = 0 \\
D_i &= 1 \quad \text{if } D_i^* > 0; \quad D_i = 0 \quad \text{otherwise} \\
D_i^* &= Z_i\delta - W_i\eta + u_i - v_i
\end{aligned}
$$

Define further

$$
s_i = \frac{-Z_i\delta + W_i\eta}{\sigma_{u-v}}
$$

The inverse Mills' ratio is

$$
\lambda_i = \frac{f(s_i)}{1 - F(s_i)}
$$

Marginal Treatment Effects: Bjorklund & Moffitt (1987)

Based on **truncated normal distributions**, one can quantify several **economically interesting effects**

Expected gain from treatment for those who select into treatment

$$\begin{aligned} E\left(\alpha_i \mid D_i\right. &= 1, Z_i\delta, W_i\eta) \\ &= Z_i\delta + E\left(u_i \mid u_i - v_i > -Z_i\delta + W_i\eta\right) \\ &= Z_i\delta + \left(\sigma_{u,u-v}/\sigma_{u-v}\right)\lambda_i \end{aligned}$$

Change in the expected gain from a change in costs

$$\partial \frac{E\left(\alpha_i \mid T_i = 1, Z_i\delta, W_i\eta\right)}{\partial\left(W_i\eta\right)} = \left[\left(\sigma_{u,u-v}\right)/\sigma_{u-v}^2\right]\lambda_i\left(\lambda_i - s_i\right) > 0$$

$\Rightarrow$ a reduction in costs *lowers* the average returns to treatment

Marginal Treatment Effects: Bjorklund & Moffitt (1987)

They derive similar effects for the outcome, $E(Y_i|\dots)$.

Key **insights from the model**:

- An instrument affects **selection into treatment at the margin**
- The estimated effects depend on **who the instrument shifts into or out of treatment**
- With **parametric assumptions**, marginal treatment effects can be estimated

Back to the Roy Model

Basic set-up:

$$
\begin{aligned}
Y_{i1} &= \mu_1 + U_{i1} \\
Y_{i0} &= \mu_0 + U_{i0} \\
D^* &= \alpha + \beta Z_i + V_i \\
E(U_{ij}) &= 0
\end{aligned}
$$

Individual treatment effect

$$
\Delta_i = Y_{i1} - Y_{i0} = \underbrace{\mu_1 - \mu_0}_{\text{common gain}} + \underbrace{U_{i1} - U_{i0}}_{\text{idiosyncratic gain}}
$$

Using this framework, we will revisit the question "**What parameter does (and should) our estimator identify?**"

LATE and other Parameters Revisited

$$ATE(x) \equiv E[Y_1 - Y_0|X = x] = \mu_1(x) - \mu_0(x)$$

$$\begin{aligned} ATT(x) &\equiv E[Y_1 - Y_0|X = x, D = 1] \\ &= \mu_1(x) - \mu_0(x) + E[U_1 - U_0|X = x, D = 1] \end{aligned}$$

$$\begin{aligned} ATU(x) &\equiv E[Y_1 - Y_0|X = x, D = 0] \\ &= \mu_1(x) - \mu_0(x) + E[U_1 - U_0|X = x, D = 0] \end{aligned}$$

$$\begin{aligned} LATE(x) &\equiv E[Y_1 - Y_0|X = x, D_1 > D_0] \\ &= \mu_1(x) - \mu_0(x) + E[U_1 - U_0|X = x, D_1 > D_0] \end{aligned}$$

How Relevant are ATE, ATT and ATU?

Even **ATE, ATT and ATU** may not be the most **policy-relevant parameters**

- They assume that individuals **cannot choose whether to take the treatment**
- In most cases this is neither feasible nor interesting

ATE: average effect if ***all*** **individuals were forced to take the treatment**

ATT: average loss for the **treated group** when switching from a regime with **optional treatment to no treatment**

ATU: average gain in the **control group if treatment were made mandatory**

The Policy-Relevant Treatment Effect

Idea goes back to Heckman & Vytlacil (2001a): consider a **policy change** that

- affects the propensity score $P(X_i, Z_i)$
- without affecting potential outcomes (Y_{i0}, Y_{i1}) or unobserved selection V_i

$\Rightarrow$ changes **who selects into treatment**

Treatment choices:

- D_i treatment choice under **baseline policy**
- $\widetilde{D}_i$ treatment choice under **alternative policy**

The Policy-Relevant Treatment Effect

PRTE: **Mean effect** of going from a **baseline to an alternative policy** per **net person shifted**

$$
\begin{aligned}
\text{PRTE}(x) &= \frac{E\left[Y_i | X_i = x, \text{ alternative policy }\right] - E\left[Y_i | X_i = x, \text{ baseline policy }\right]}{E\left[D_i | X_i = x, \text{ alternative policy }\right] - E\left[D_i | X_i = x, \text{ baseline policy }\right]} \\
&= \mu_1(x) - \mu_0(x) \\
&+ \frac{E\left[U_{1i} - U_{0i} | X_i = x, \tilde{D}_i = 1\right] E\left[\tilde{D}_i | X_i = x\right] - E\left[U_{1i} - U_{0i} | X_i = x, D_i = 1\right] E\left[D_i | X_i = x\right]}{E\left[\tilde{D}_i\right] X_i = x] - E\left[D_i | X_i = x\right]}
\end{aligned}
$$

LATE and other Parameters

LATE coincides with ATE, ATT, ATU and PRTE if one of two conditions is met

1) If **treatment effects are homogeneous**

2) If treatment effects are heterogeneous but individuals **do not select into treatment based on their treatment effects**

- **unrealistic outside experiments** (and even within if there is imperfect compliance)
- Those with the **largest gains should be most likely to comply**

Marginal Treatment Effects (MTE)

How to get **from LATE to ATE or PRTE**?

- Need to **extrapolate to always- and never-takers**
- This requires **making strong assumptions**

Enter **Marginal Treatment Effects**....

- approach allows to **recover policy-relevant parameters**
- by relying on **continuous** or at least **non-binary instruments**

Marginal Treatment Effects (MTE)

MTE explicitly **models the selection into treatment**

It is based on **potential outcomes** (similar to LATE)

Potential outcome equations are decomposed into

- an **observed component**
- and an **unobserved component**

The "**MTE Curve**" relates

- the **unobserved heterogeneity in the treatment effect**
- to the unobserved heterogeneity in the **propensity of taking the treatment**

The Marginal Treatment Effect (MTE)

For given v and x, the **Marginal Treatment Effect** is

$$MTE(v, x) \equiv E[Y_1 - Y_0 | V = v, X = x]$$

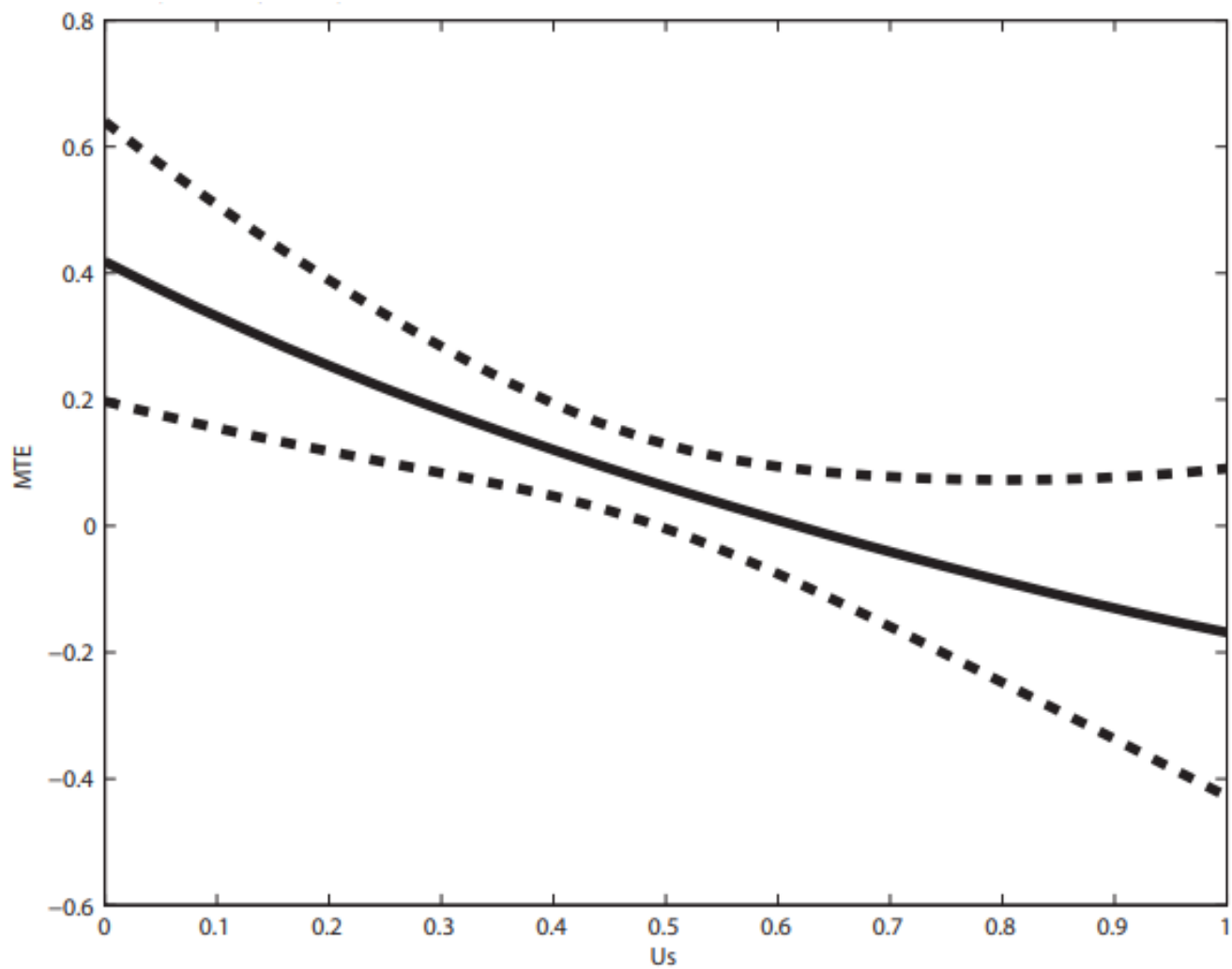
The MTE is a **function of unobservable** and **observable** characteristics

It allows for **heterogeneity in treatment effects** along both dimensions

Usually **MTE declines with v**: treatment is chosen by people with the largest gains

Our Goal: the MTE Curve

Example from Carneiro *et al.* (2011): returns to education



Higher U: higher resistance to treatment

Marginal Treatment Responses

Mogstad & Torgovitsky (2018) split the MTE in separate **marginal treatment responses (MTR)** for the treated and control group

$$m_0(v, x) \equiv E[Y_0|V = v, X = x] \quad \text{and} \quad m_1(v, x) \equiv E[Y_1|V = v, X = x]$$

Advantage of MTRs:

- many **interesting parameters** can be re-written as **weighted averages of m_0 and m_1**
- the weights are often **asymmetric**

Numerical Example

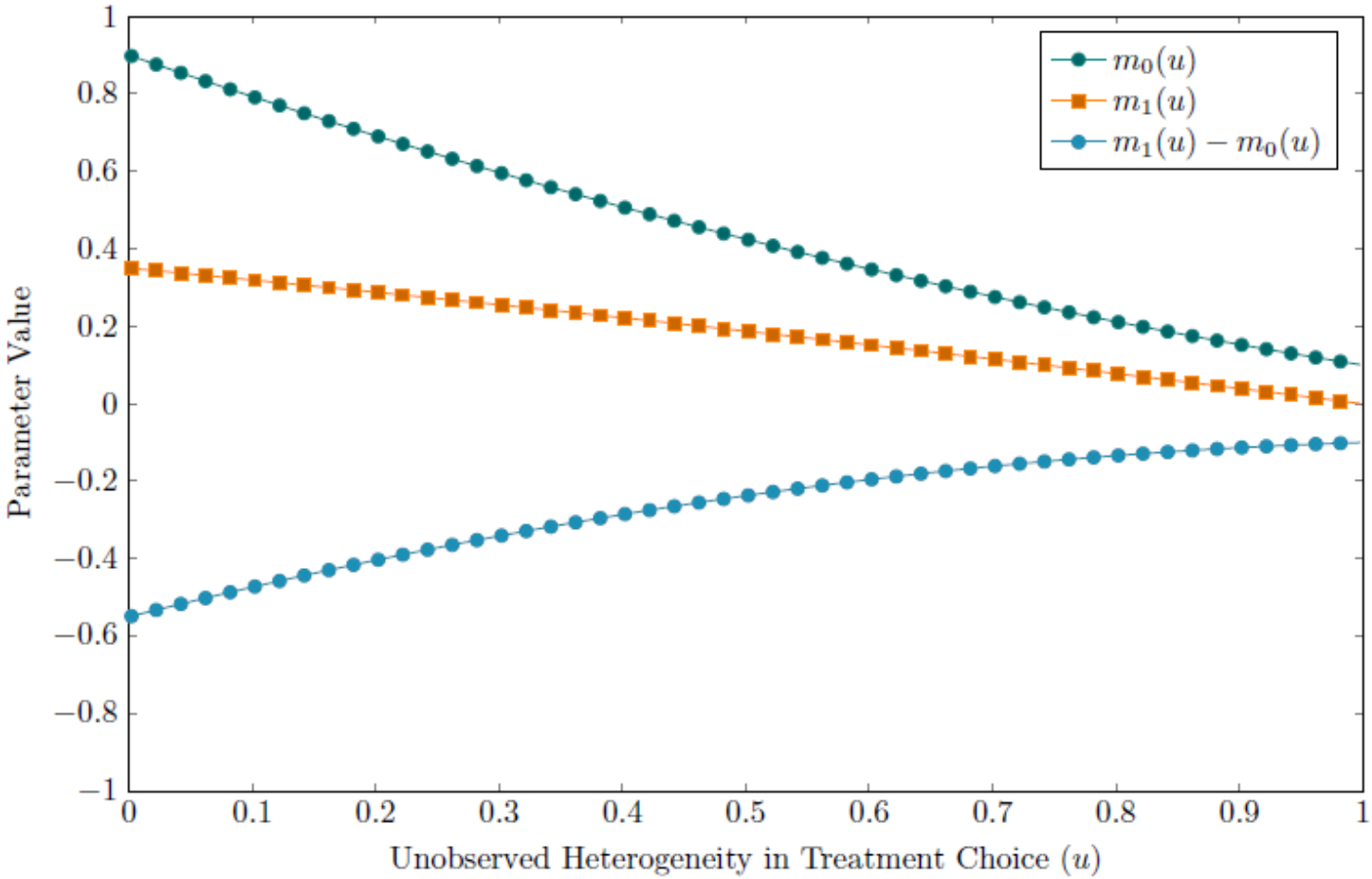
From Mogstad & Torgovitsky (2018): **effect of mosquito nets (D) on malaria (Y)**

- Z is a **randomly assigned subsidy** with different levels $Z \in \{1, 2, 3, 4\}$
- propensity scores $p(z) \in \{0.12, 0.29, 0.48, 0.78\}$

Next slide shows:

- malaria infections are **lower in the treatment group**
- people with the **biggest gains are most likely to buy a net** $\Rightarrow$ **heterogeneous treatment effects**

Numerical Example



The Target Parameter

The **target parameter $\beta^\star$ for any policy** can be expressed as the **weighted average of the two unknown MTR functions**

$$\beta^\star \equiv E\left[\int_0^1 m_0(v,X)\omega_0^\star(v,X,Z)dv\right]+E\left[\int_0^1 m_1(v,X)\omega_1^\star(v,X,Z)dv\right]$$

$\omega_0^\star$ and $\omega_1^\star$ are **weighting functions**

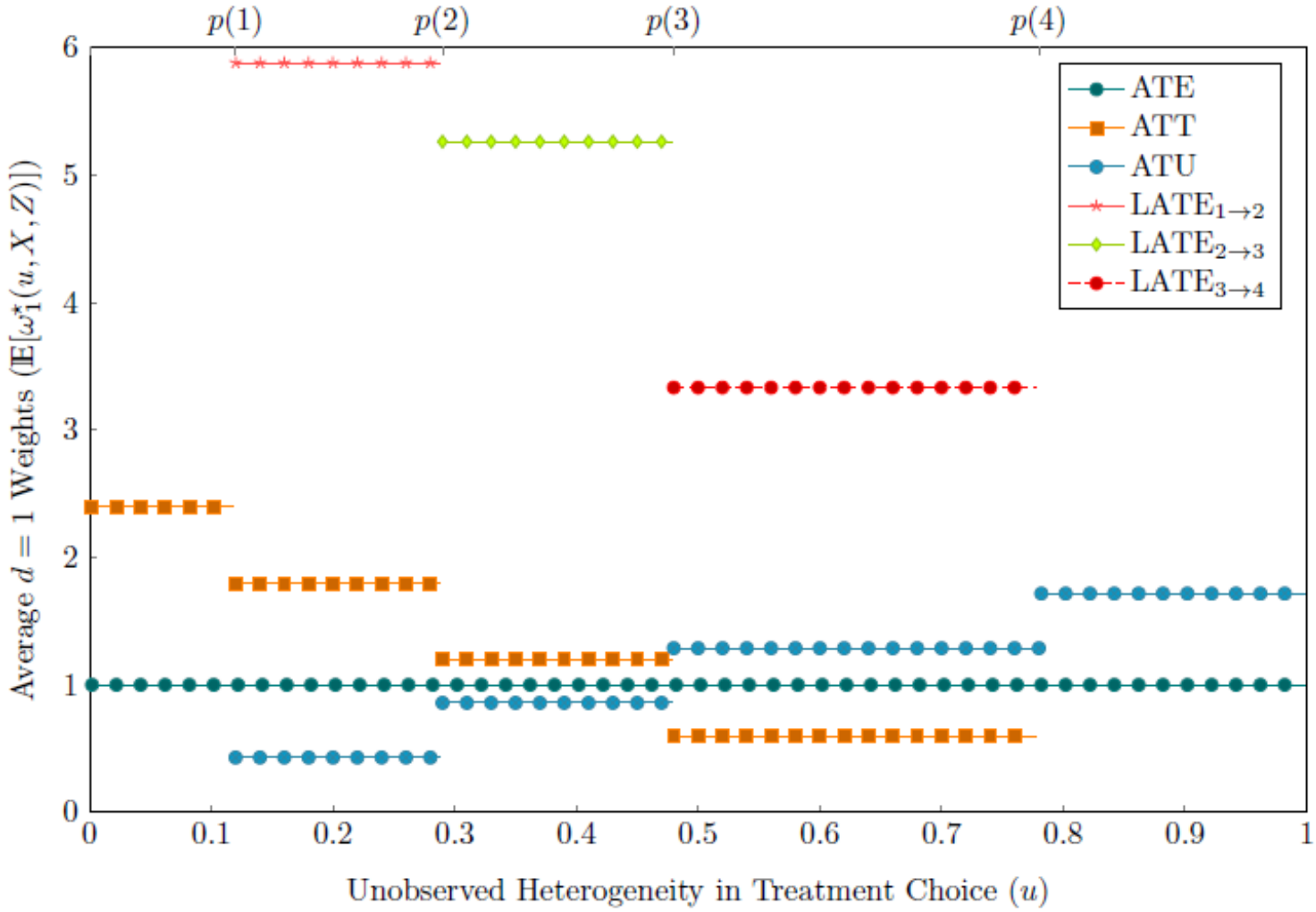
- **weight of each observation** in the MTR
- these need to be **chosen or estimated**

Weighting Functions for Common Parameters

Target Parameter	Expression	Weights $\omega_0^\star(u,x,z)$	Weights $\omega_1^\star(u,x,z)$
Average Untreated Outcome	$E[Y_0]$	1	0
Average Treated Outcome	$E[Y_1]$	0	1
Average Treatment Effect (ATE)	$E[Y_1 - Y_0]$	-1	1
ATE given $X = \overline{x}$ where $P[X = \overline{x}] > 0$	$E[Y_1 - Y_0 \mid X = \overline{x}]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[x = \overline{x}]}{P[X = \overline{x}]}$
Average Treatment on the Treated (ATT)	$E[Y_1 - Y_0 \mid D = 1]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[u \leq p(x,z)]}{P[D = 1]}$
Average Treatment on the Untreated (ATU)	$E[Y_1 - Y_0 \mid D = 0]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[u > p(x,z)]}{P[D = 0]}$
Local Average Treatment Effect (LATE) for $z \to z'$ given $X = x$, where $p(x,z') > p(x,z)$	$E[Y_1 - Y_0 \mid p(x,z) < U \leq p(x,z'), X = x]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[p(x,z) < u \leq p(x,z')]}{p(x,z') - p(x,z)}$

From Mogstad & Torgovitsky (2018). Note: they use *u* instead of *v*. More intuition on the next slide

Weighting Functions for Common Parameters



⇒: ATE, ATT and ATU differ in their weights, as do LATEs

Identification Assumptions Revisited

Independence of the Instrument

$$(U_0, U_1, V) \perp Z|X$$

Standard LATE assumptions: **strong first stage, monotonicity**

No need to assume:

- independence of V and U_j; unobserved returns $(U_1 - U_0)$ and resistance to treatment can be correlated
- independence of X and U_j

Propensity to be Treated

Let F_V be the CDF of V

Propensity score:

$$P(z) = \Pr(D = 1|Z = z, X = x) = F_V\left(\mu_D(X, Z)\right)$$

Convention in the MTE literature: define $U_D = F_V(V)$

- U_D is the **quantile** of the resistance distribution
- U_D is uniformly distributed

Treatment choice depends on the instrument $\Rightarrow$ observed encouragement exceeds unobserved resistance

$$D = 1 \text{ if } P(z) \geq U_D$$

Back to LATE and a Binary Instrument

For a given value $X = x$ and a binary instrument, the **Wald Estimator** is

$$\text{Wald}\,(x) = \frac{E[Y|Z=1, X=x] - E[Y|Z=0, X=x]}{E[D|Z=1, X=x] - E[D|Z=0, X=x]}$$

And the **LATE** is

$$\begin{aligned}\text{LATE}(x) = \quad & E\left[Y_1 - Y_0 | D_1 > D_0, X = x\right] \\ = & \mu_1(X) - \mu_0(X) + E\left[U_1 - U_0 | D_1 > D_0, X = x\right]\end{aligned}$$

$\Rightarrow$ effect of treatment on the compliers

LATE with a Continuous Instrument

MTE becomes a lot more insightful with a **continuous instrument**. Think about it as **pairwise comparisons** of values of Z

Consider a pair of values z and z'. The pairwise **Wald Estimator** becomes

$$\text{Wald } (z, z', x) = \frac{E\left[Y_i | Z_i = z, X_i = x\right] - E\left[Y_i | Z_i = z', X_i = x\right]}{E\left[D_i | Z_i = z, X_i = x\right] - E\left[D_i | Z_i = z', X_i = x\right]}$$

Assume that moving from z' to z **shifts units into treatment**

$$E[D|Z = Z, X = x] > E\left[D|Z = z', X = x\right]$$

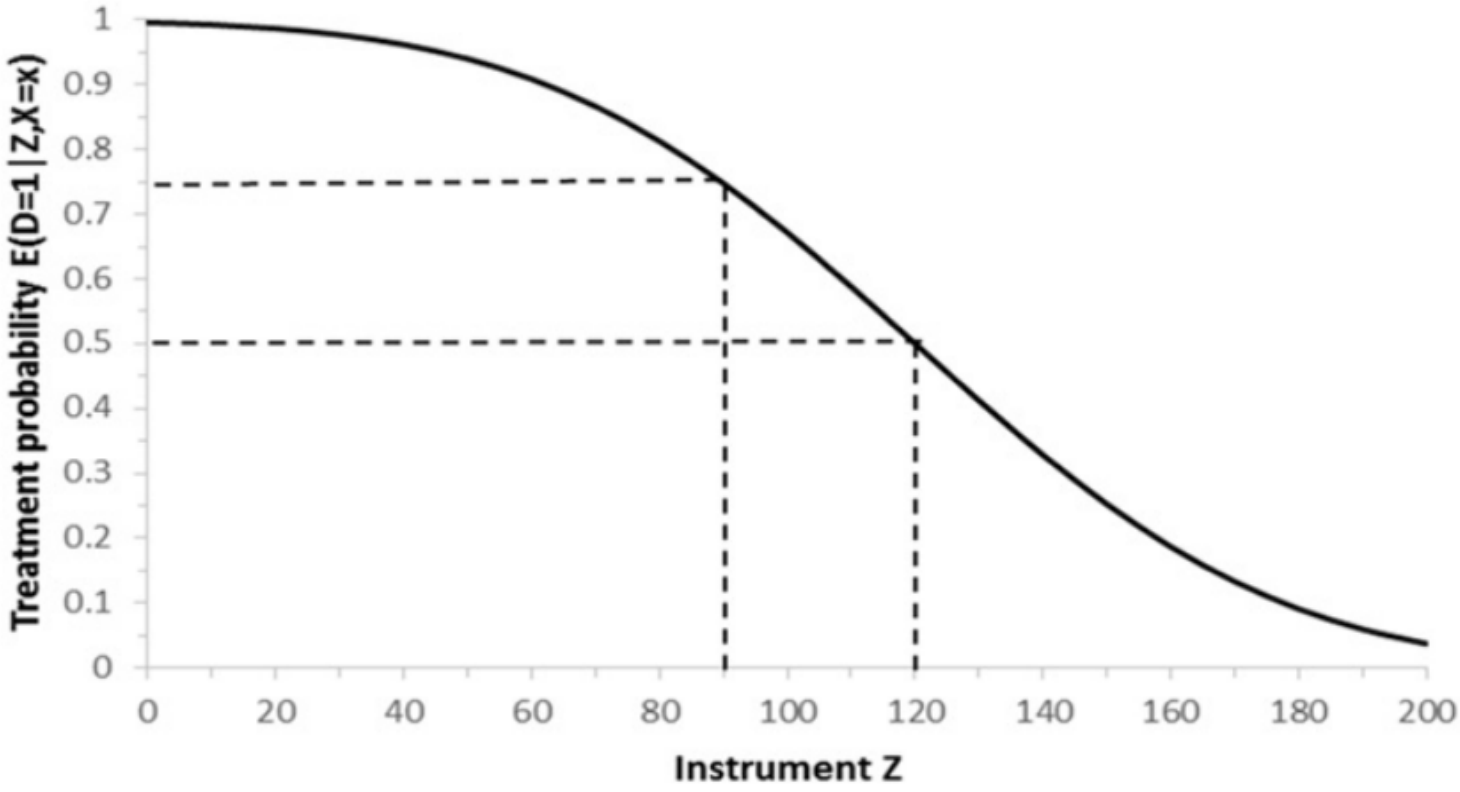
The **pairwise LATE** for given x is

$$\text{LATE}\left(z, z', x\right) = E\left[Y_1 - Y_0 | P\left(z'\right) < \mathrm{U_D} < P(z), X = x\right]$$

$\Rightarrow$ ATE for those **complying with the change of the instrument**

Compliers with a Continuous Instrument

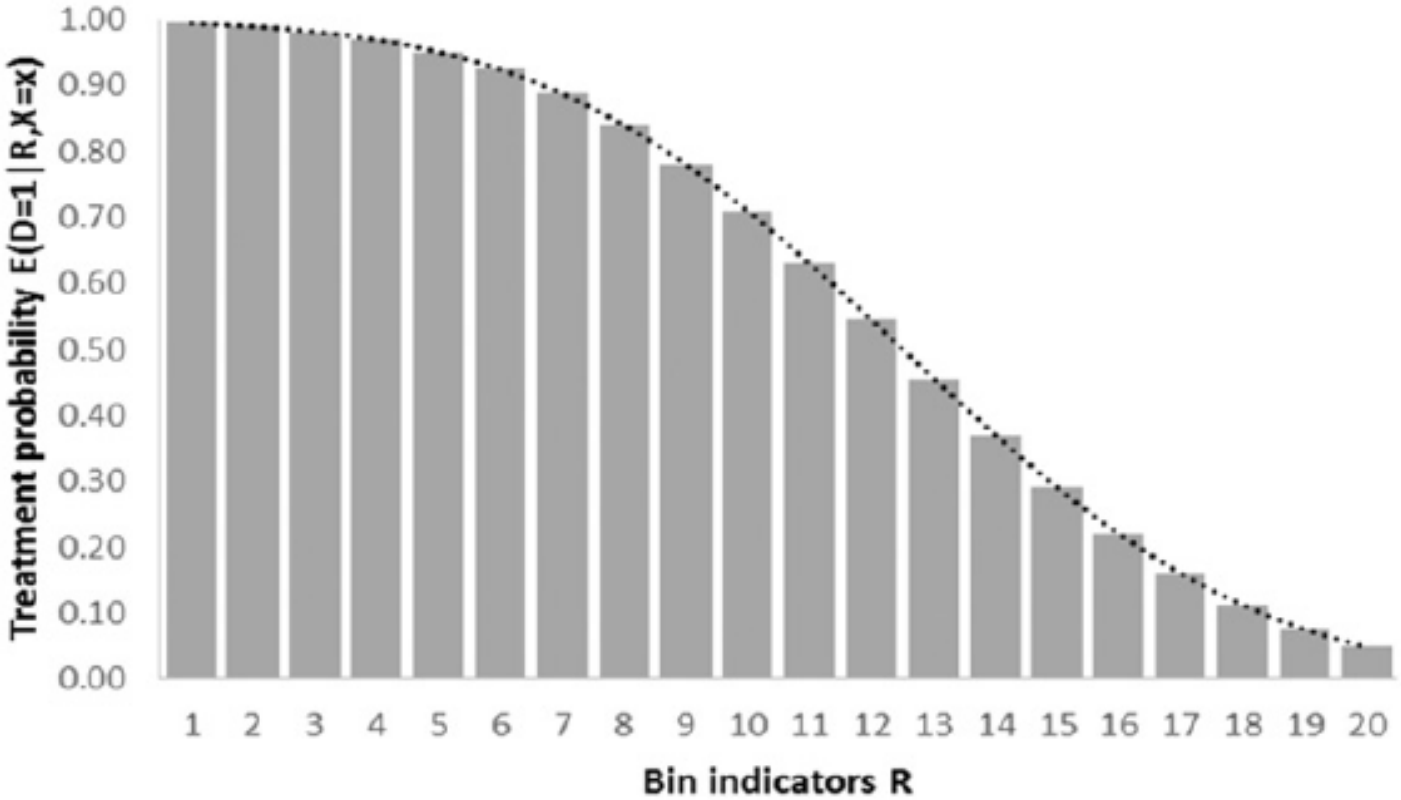
Example from Cornelissen *et al.* (2016) (Z: distance to college)



Decreasing Z from 120 to 90 **shifts individuals with** $0.5 < U_D < 0.75$ **into treatment**

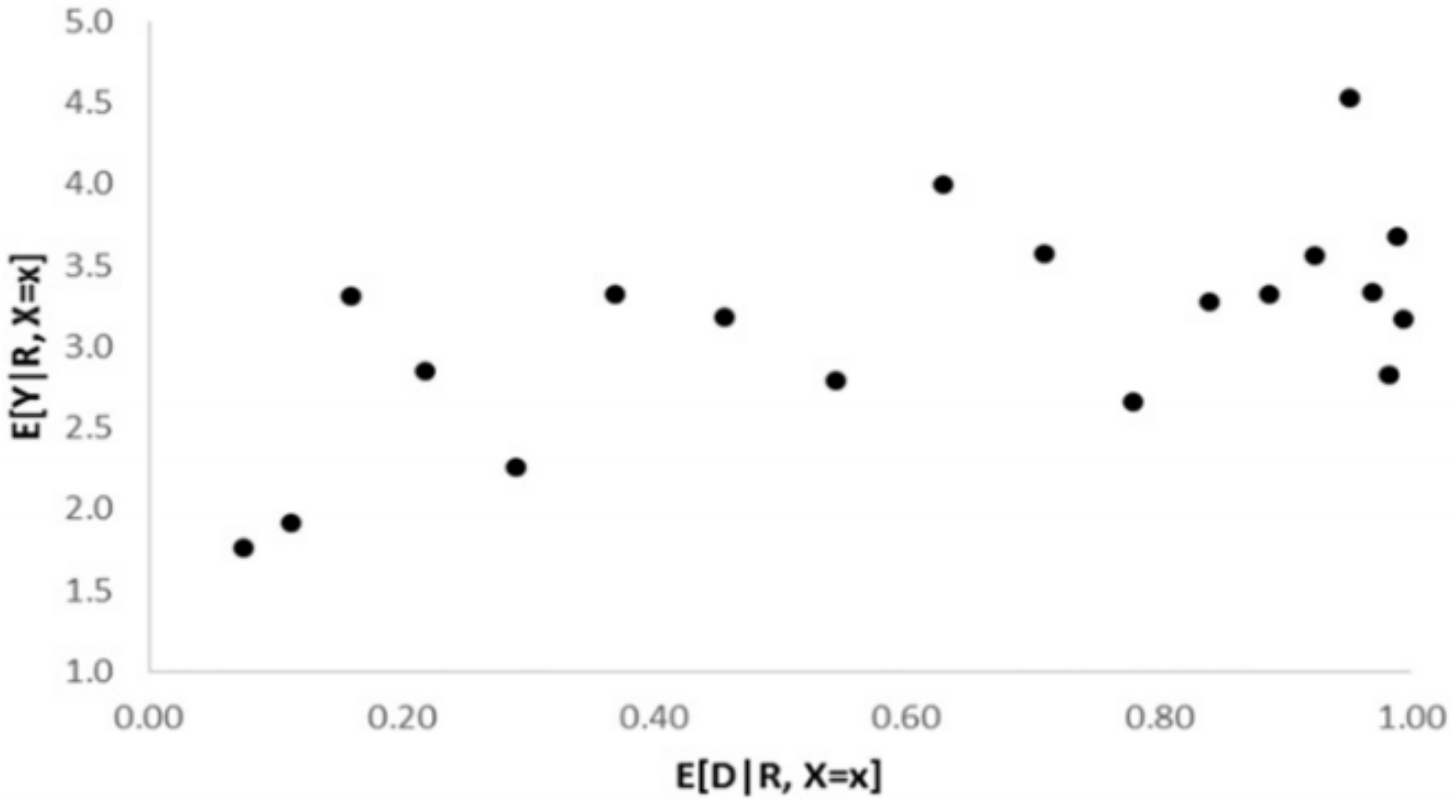
We Can't Compare all Values of z

Sol: use a **finite number of bins**



2SLS with a Continuous Instrument

To obtain the LATE for $X = x$, we fit a line through this cloud of points...



Source: Cornelissen *et al.* (2016)

2SLS with a Continuous Instrument

To obtain the **overall IV estimator**, we obtain the LATE for every value of $X = x$

The IV estimator is the **variance-weighted average of all LATEs**

$$IV = \sum_{x \in \mathcal{X}} \omega(x) \, \text{LATE}(x)$$

Weights are equal to the contribution of units with $X = x$ to the first stage

Back to the Marginal Treatment Effect

$$\begin{aligned} MTE\left(X, u_D\right) &= E\left(Y_1 - Y_0 | X = x, U_D = u_D\right) \\ &= \mu_1\left(X_i\right) - \mu_0\left(X_i\right) + E\left(U_1 - U_0 | X = x, U_D = u_D\right) \end{aligned}$$

Interpretation: **treatment effect of a person**

- with observable characteristics x
- who is at the u_D^{th} quantile of the distribution of V

An individual with $u_D = 0.1$ and a propensity score $P(Z = z) = 0.1$ is **indifferent between taking the treatment or not**

The Marginal Treatment Effect

For a given X and $P(z)$, the MTE is given by

$$\text{MTE}\left(X = x, U_D = p\right) = \frac{\partial \text{E}(Y|X = x, P(Z) = p)}{\partial p}$$

- Let individuals with $U_D = p_0$ be **indifferent to taking the treatment or not**
- Increasing p by a small amount dp **shifts them into treatment**
- Their **marginal treatment effect** at p_0 is $MTE(U_D = p_0)$
- The **increase in the average outcome** for that group is $dY = dp \times MTE(U_D = p_0)$

Therefore: $\frac{dY}{dp} = MTE(U_D = p_0)$

How to Estimate the MTE

Challenge: we don't know U_1, U_0, V

Solution 1: **parametric estimation** (Bjorklund & Moffitt, 1987)

- assume $(U_1, U_0, V) \sim N(0, \mathbf{\Sigma})$
- estimate with ML; assumptions too strong

Solution 2: **fully non-parametric** (Heckman & Vytlacil, 1999, 2001b, 2005)

- separate flexible function $Y(P(Z))$ for each group defined by $X = x$
- very demanding on the data

Solution 3: **shape restrictions** (Cornelissen *et al.*, 2016)

- linear separable potential outcomes $Y_j = X_j\beta_j + U_j$
- MTE curve independent of X

Estimating MTE under Linear Separability and Shape Restrictions

Under these assumptions, the **MTE becomes additively separable**

$$\begin{aligned} \text{MTE}(x, u_D) &= E(Y_{1i} - Y_{0i} | X_i = x, U_{Di} = u_D) \\ &= \underbrace{x(\beta_1 - \beta_0)}_{\text{observed component}} + \underbrace{E(U_{1i} - U_{0i} | U_{Di} = u_D)}_{\text{unobserved component}} \end{aligned}$$

This leads to the **outcome equation**

$$E[Y_i | X_i = x, P(Z) = p] = X_i\beta_0 + X_i(\beta_1 - \beta_0)p + K(p)$$

$K(p)$ is a non-linear function of the propensity score

Estimating MTE under Linear Separability and Shape Restrictions

The Marginal Treatment Effect is then given by

$$\text{MTE}\left(X_i = x, U_{Di} = p\right) = \frac{\partial E\left[Y_i | X_i = x, P(Z) = p\right]}{\partial p} = x\left(\beta_1 - \beta_0\right) + \frac{\partial K(p)}{\partial p}$$

What to do with $K(p)$?

- model as polynomial of p
- semi- or non-parametric modelling

How to Estimate the MTE

1) Estimate the **propensity score** $\hat{p}$ with probit or logit

2) Model Y as a function of X, $X\hat{p}$ and $\hat{p}$

Example from Cornelissen *et al.* (2018)

$$Y = X\beta_0 + R\alpha + T\tau + X\left(\beta_1 - \beta_0\right)\hat{p} + \sum_{k=2}^{K} \alpha_k \hat{p}^k + \varepsilon$$

3) Calculate the derivative of $\hat{Y}$ wrt $\hat{p}$ at given values of X

4) plot the MTE curve

Packages: *locallV* in R, *margte* in Stata

Example: Marginal Gains to Early Childcare

Cornelissen *et al.* (2018) study the effect of **early childcare attendance** on **school readiness**

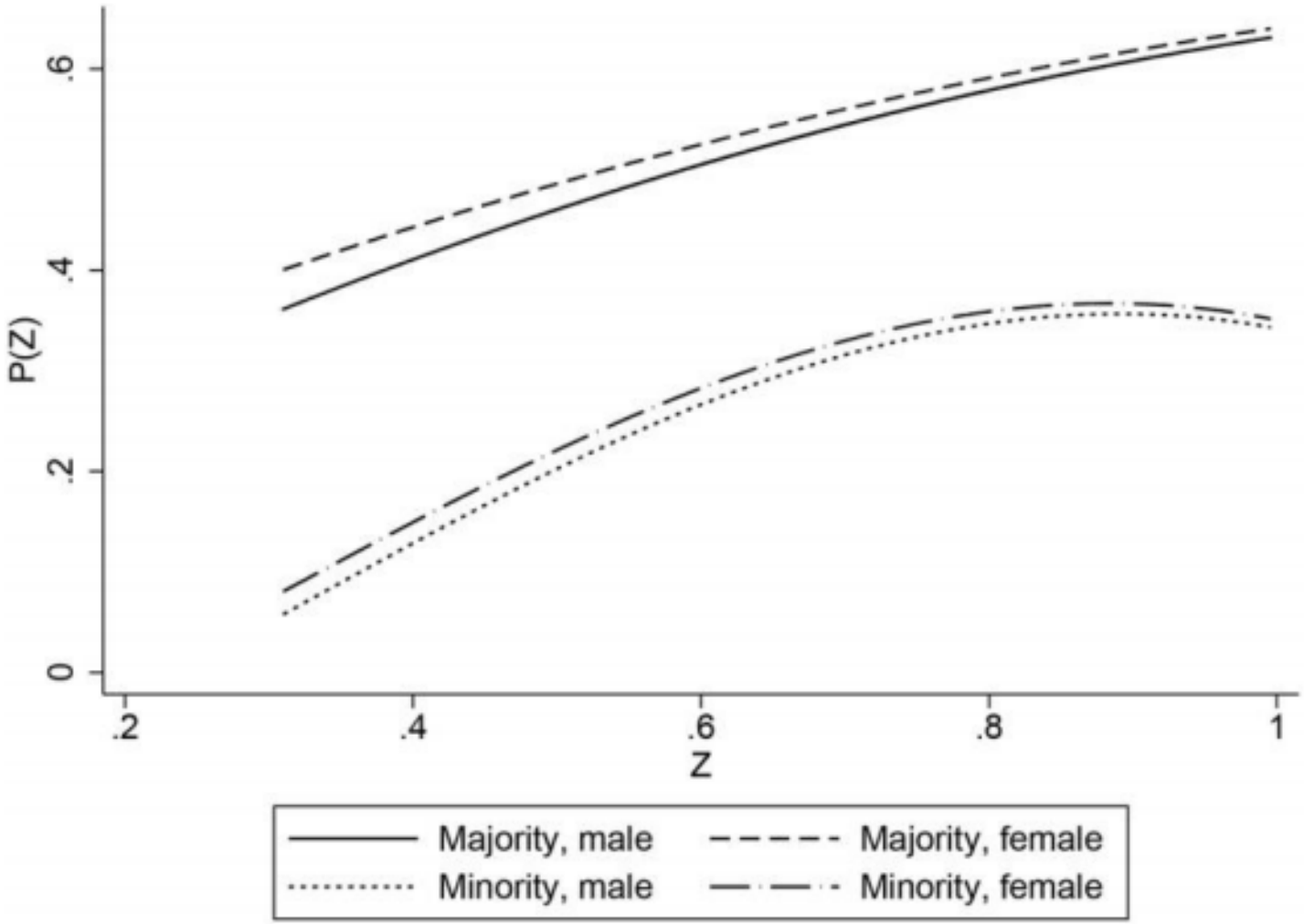
They exploit the **staggered introduction of free childcare** (age 3-6) in Germany

Instrument is the **number of available childcare slots** before age 3

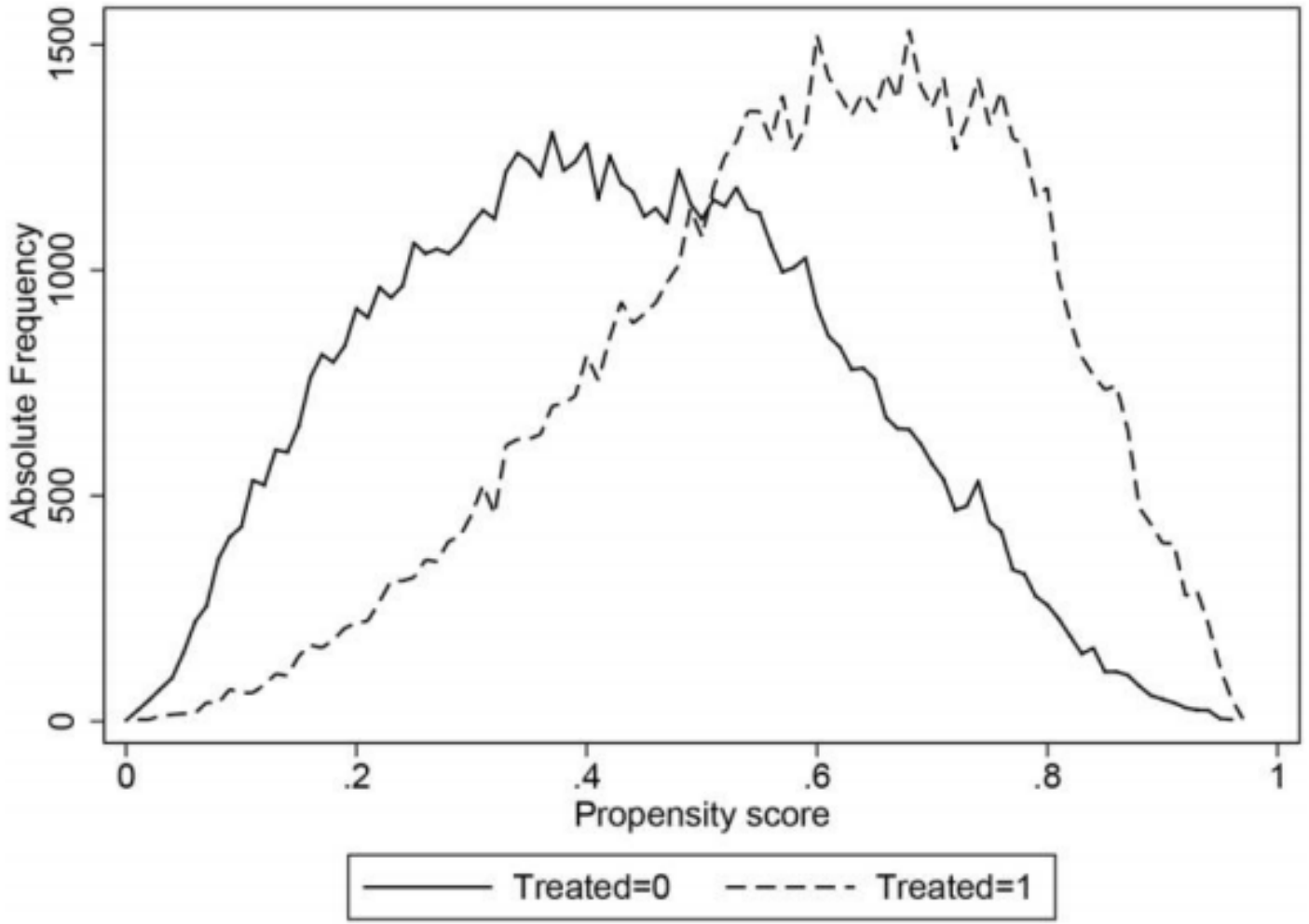
Questions asked with MTE:

- Which children are most likely to **select into childcare**?
- Which children have the **highest causal effects**?

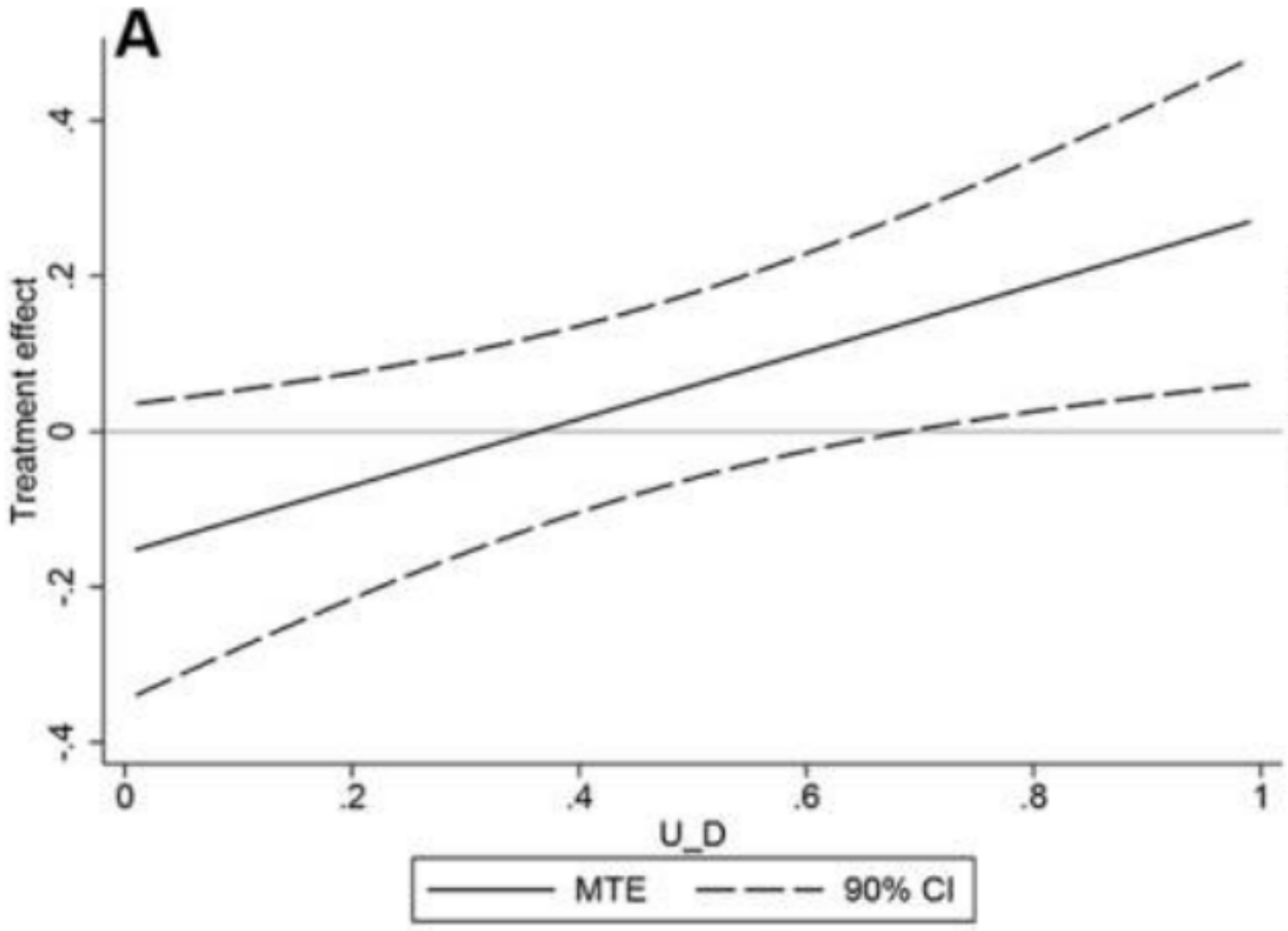
Selection into Treatment



Common Support

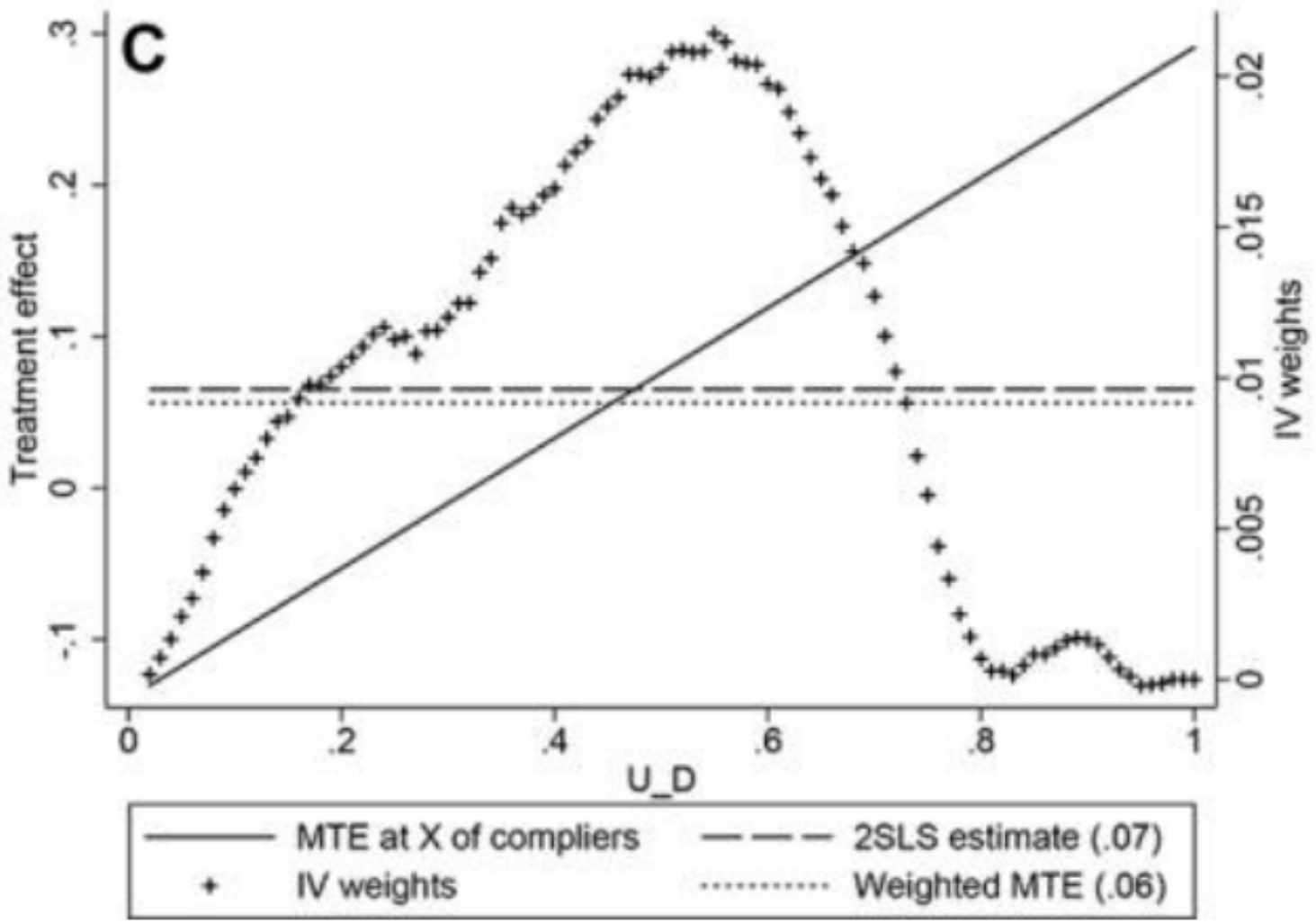


The MTE Curve



Higher U_D (resistance): lower probability of treatment

Compare with 2SLS



Lessons from Cornelissen *et al.* (2018)

Children with **small gains select into childcare**, those with high gains don't

Gains are large for those **least likely to attend childcare**

An **IV estimator** would have given **no weight to children with very high vs. very low** likelihood of attending childcare

Extrapo-LATE-ing

MTE requires an instrument with broad support

- Most instruments do not fall into this category
- They have some support, but they don't have the potential to shift everyone into treatment

Mogstad *et al.* (2018) show how MTE can be used to **extrapolate from the LATE to support not covered by the instrument**

Idea:

- Make assumptions about **MTEs outside the support** $[\underline{u}, \bar{u}]$
- Construct **bounds** on the target parameter
- Bounds are **tighter the closer** the support to $[\underline{u}, \bar{u}]$

For an application, see Brinch *et al.* (2017)

Summary: MTE

MTE is **useful and informative because**

- It explicitly **models treatment choice**
- It characterizes **heterogeneous treatment effects**
- It allows for the estimation of **policy-relevant parameters**

Downsides:

- Requires a large dataset
- Requires a credible instrument with broad support

More advanced methods are summarized in Mogstad & Torgovitsky (2018)

References I

Bjorklund, Anders, & Moffitt, Robert. 1987. The Estimation of Wage Gains and Welfare Gains in Self-selection. *The Review of Economics and Statistics*, **69**(1), 42–49.

Brinch, Christian N., Mogstad, Magne, & Wiswall, Matthew. 2017. Beyond LATE with a Discrete Instrument. *Journal of Political Economy*, **125**(4), 985–1039.

Carneiro, Pedro, Heckman, James J., & Vytlacil, Edward J. 2011. Estimating Marginal Returns to Education. *American Economic Review*, **101**(6), 2754–81.

Cornelissen, Thomas, Dustmann, Christian, Raute, Anna, & Schönberg, Uta. 2016. From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, **41**(C), 47–60.

Cornelissen, Thomas, Dustmann, Christian, Raute, Anna, & Schönberg, Uta. 2018. Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance. *Journal of Political Economy*, **126**(6), 2356–2409.

Heckman, James J. 1978. Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, **46**(4), 931–959.

Heckman, James J. 1979. Sample Selection Bias as a Specification Error. *Econometrica*, **47**, 153–161.

Heckman, James J., & Vytlacil, Edward. 2001a. Policy-Relevant Treatment Effects. *The American Economic Review, Papers & Proceedings*, **91**(2), 107–111.

Heckman, James J., & Vytlacil, Edward. 2005. Structural Equations, Treatment Effects, and Econometric Policy Evaluation1. *Econometrica*, **73**(3), 669–738.

Heckman, James J., & Vytlacil, Edward J. 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, **96**(8), 4730–4734.

Heckman, James J., & Vytlacil, Edward J. 2001b. *Local instrumental variables*. International Symposia in Economic Theory and Econometrics. Cambridge University Press. Page 1–46.

Mogstad, Magne, & Torgovitsky, Alexander. 2018. Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics*, **10**(1), 577–613.

Mogstad, Magne, Santos, Andres, & Torgovitsky, Alexander. 2018. Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters. *Econometrica*, **86**(5), 1589–1619.