Marginal Treatment Effects (MTE)

ECOM113 Advanced Topics in Microeconomics

Lecturer: Anna Raute

(Slides are complimentary to Cornelissen et al. (2016))

MTE: some key literature

- **Well-known introduction/application:** Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2011. “Estimating Marginal Returns to Education.” *American Economic Review* 101(6): 2754–2781.

- **One of the first papers:** Björklund, Anders, and Robert Moffitt. 1987. “The Estimation of Wage Gains and Welfare Gains in Self-Selection Models.” *Review of Economics and Statistics* 69(1): 42–49.

- **Practical applied:**

  - Brave, Scott, and T. Walstrum. 2014. “Estimating Marginal Treatment Effects Using Parametric and Semiparametric Methods.” *Stata Journal* 14(1): 191–217.

  - Cornelissen, T., Dustmann, C., Raute, A, and U. Schönberg 2016: From LATE to MTE: Alternative Methods for the Evaluation of Policy Interventions, *Labour Economics*, 41, 47-60.

  - Andresen, Martin. 2017. “Exploring Marginal Treatment Effects: Flexible estimation using Stata”, available at https://sites.google.com/site/martineckhoffandresen

MTE: some key literature

- **More technical:** Heckman, James J, Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics* 88(3): 389–432.

- **Very comprehensive:** Heckman, James J., and Edward J. Vytlacil. 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." Chap. 71 in *Handbook of Econometrics*, ed. by James J. Heckman and Edward E. Leamer. Amsterdam: Elsevier.

Motivation of MTE

- Recall that with heterogeneous treatment effects, IV estimates a local average treatment effect (LATE), a causal effect for the group of compliers.
- LATE is entirely defined by the instrument that is being used, not by an economic policy question. If compliers are a peculiar group, then LATE may not be interesting and may not answer an economic policy question.
- The MTE framework aims at using instruments in a way that allows recovering economically interesting parameters. To avoid overly strong assumptions, one needs continuous instruments!

Typical economically interesting parameters

- What are economically interesting parameters?
  - Average Treatment Effect on the whole population (ATE) (also interpretable as effect on a person picked at random from the population)
  - Average Treatment Effect on the Treated (ATT)
    Average Treatment Effect on the Untreated (ATU)
  - Average Treatment Effect on a subpopulation that would be shifted into treatment by a specific policy change (Policy-relevant treatment effect, PRTE)
- In some special cases LATE coincides with these economically interesting parameters

Special cases

- Cases in which LATE, ATE, ATT, ATU are the same:
  - if treatment effects are homogeneous (do not vary across individuals)
  - or if treatment effects are heterogeneous, but individuals do not select into treatment based on (the unobserved component of) their individual treatment response

The latter “is a strong assumption that forces the analyst to assume either irrationality or ignorance on the part of persons whose behavior is being studied” (Heckman 1997, JHR)

Marginal Treatment Effects (MTE)

Ingredients of the MTE framework:

- Potential outcome equations in the treated and untreated state (similar to the LATE framework).
- Distinguish observed and unobserved part of the potential outcome equations, and thus an observed and unobserved part of the treatment effect.
- Latent index model for treatment selection, with observed and unobserved determinants of selection into treatment.

Allows to relate the unobserved heterogeneity in the treatment effect to the unobserved heterogeneity in the propensity of taking the treatment (“MTE curve”).

Potential Outcome Equations

Potential outcome for treated (Y_1) and untreated (Y_0):

$$Y_1 = \mu_1(X) + U_1$$
$$Y_0 = \mu_0(X) + U_0$$

with $\mu_j(X) = E\left(Y_j | X = x\right)$. If one wanted to assume linearity this could be modelled as $\mu_j(X) = X\beta_j$, for $j = 0,1$.

Heckman and coauthors also discuss the more general case of nonseparable errors, i.e., $Y_j = \mu_j\left(X, U_j\right)$ for $j = 0,1$.

Define a treatment indicator with $D = 1$ if an individual is treated and $D = 0$ otherwise.

Potential Outcome Equations

The observed outcome (Y) is then modelled as:

$$Y = DY_1 + (1 - D)Y_0 = Y_0 + \underbrace{(Y_1 - Y_0)}_{\boldsymbol{\Delta}} D,$$

where

$$\boldsymbol{\Delta} = Y_1 - Y_0 = \underbrace{\mu_1(X) - \mu_0(X)}_{\text{observed}} + \underbrace{U_1 - U_0}_{\text{unobserved}}$$

is the treatment effect, which consists of an observed and unobserved part. For the parameterisation $\mu_j(X) = X\beta_j$, for $j = 0{,}1$, the observed part simplifies to:

$$\mu_1(X) - \mu_0(X) = X\beta_1 - X\beta_0 = X(\beta_1 - \beta_0)$$

Treatment choice

Choice modelled by an index threshold crossing model with latent net benefit of choosing the treatment D^* equal to:

$$D^* = \mu_D(X, Z) - V$$

X & Z: observed determinants of treatment choice, where Z is excluded from the outcome equation

V: unobserved characteristics that make treatment choice less likely (unobserved “resistance” to take treatment)

Treatment is chosen if net gain of treatment is positive:

$$D = 1 \text{ if } D^* \geq 0; \quad \mathrm{D} = 0, \text{otherwise.} \qquad (1)$$

In the case of a linear index model: $\mu_D(X, Z) = X\gamma + Z\delta$.

Assumptions

- (U_0, U_1, V) is statistically independent of Z given X:
$$(U_0, U_1, V) \perp\!\!\!\perp Z \mid X$$
(The instrument Z must be as good as randomly assigned conditional on X)

- V can depend on U_1 and U_0 in a general way (meaning that unobserved returns $U_1 - U_0$ are associated with unobserved resistance V to take the treatment)

- U_1 and U_0 do not need to be statistically independent of X, but in applications some restrictions are usually assumed

- Additional assumptions (as in LATE-framework): Existence of first stage and Monotonicity (Uniformity)

The propensity score

- Let F_V be the distribution function of V.
- Define the propensity score as

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

- Define $U_D = F_V(V)$, this transforms the unobserved resistance to treatment V into its quantiles U_D. By construction U_D is uniformly distributed.
- The treatment choice decision (1) can now be written as:

$$D = 1 \quad \text{if} \quad P(z) \geq U_D, \qquad D = 0, \text{otherwise.}$$

(Treatment is chosen if “observed encouragement” exceeds “unobserved resistance”.)

Summary Treatment effects

- $ATE(x) = E[\mathbf{\Delta}|X = x] = \mu_1(x) - \mu_0(x)$
- $ATT(x) = E[\mathbf{\Delta}|X = x, D = 1] = \mu_1(x) - \mu_0(x) + E[U_1 - U_0|X = x, D_i = 1]$
- $ATU(x) = E[\mathbf{\Delta}|X = x, D = 0] = \mu_1(x) - \mu_0(x) + E[U_1 - U_0|X = x, D = 0]$

- Consider a policy change which affects the propensity score. Let D_i ($\widetilde{D}_i$) be the treatment choice under the baseline (alternative) policy

- $$PRTE(x) = \frac{E[Y|X=x,\text{alternative policy}]-E[Y|X=x,\text{baseline policy}]}{E[D|X=x,\text{alternative policy}]-E[D|X=x,\text{baseline policy}]}$$
  $$= \mu_1(x) - \mu_0(x)$$
  $$+ \frac{E[U_1-U_0|X=x,\widetilde{D}=1]E[\widetilde{D}|X_i=x]-E[U_1-U_0|X=x,D=1]E[D|X=x]}{E[\widetilde{D}|X=x]-E[D|X=x]}$$

LATE

Consider first the case of binary instrumental variable, Z.

- In this case, the IV estimator conditional on $X = x$ is equal to the Wald estimator:

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

- Let D_{1i} be i's treatment status when instrument z_i=1 and D_{0i} i's treatment status when z_i=0.
- Under standard LATE assumptions (Independence, First stage, Monotonicity/Uniformity ($D_{1i} \geq D_{0i}\ \forall i,$ i.e. no defiers ($D_{1i} = 0$ and $D_{0i} = 1$)):
- $$\begin{aligned} LATE(x) &= E[Y_1 - Y_0 | D_1 > D_0, X = x] \\ &= \mu_1(X) - \mu_0(X) + E[U_1 - U_0 | D_1 > D_0, X_i = x] \end{aligned}$$
   - LATE is effect of treatment on compliers ($D_{1i} = 1$ and $D_{0i} = 0$)
   - IV is not informative on always-takers ($D_{1i} = 1$ and $D_{0i} = 1$) or never-takers ($D_{1i} = 0$ and $D_{0i} = 0$)

Special cases

Case in which LATE is the ATT:

- One-sided non-compliance, if there are no always-takers, then LATE is ATT. Example: A randomized trial in which some people who were assigned for treatment do not take the treatment (never-takers), but nobody in the control group has access to the treatment (no always-takers).
  - » Example: Chetty et al. (2016) on long-run effects of Moving to Opportunity (randomly selected housing vouchers to move to lower-poverty area).
    - Random assignment to treatment group (offer of voucher) as instrument for actual treatment decision (relocation)
    - All treated are compliers (nobody in control group had access to treatment), hence LATE=ATT

Special cases

Case in which LATE is a policy-relevant effect (PRTE):

- If the instrument is a policy of interest, then LATE is the effect on individuals who are shifted into treatment by the policy, and thus LATE is equal to PRTE
  - » Example: Oreopoulos (2006, AER): uses increase in compulsory-school leaving age (from 14 to 15) in Britain as instrument for extra year of schooling.
    - Estimates effect for people who only stay in school longer due to policy reform and would have left at 14
    - no never-takers due to full enforcement, hence LATE=ATU

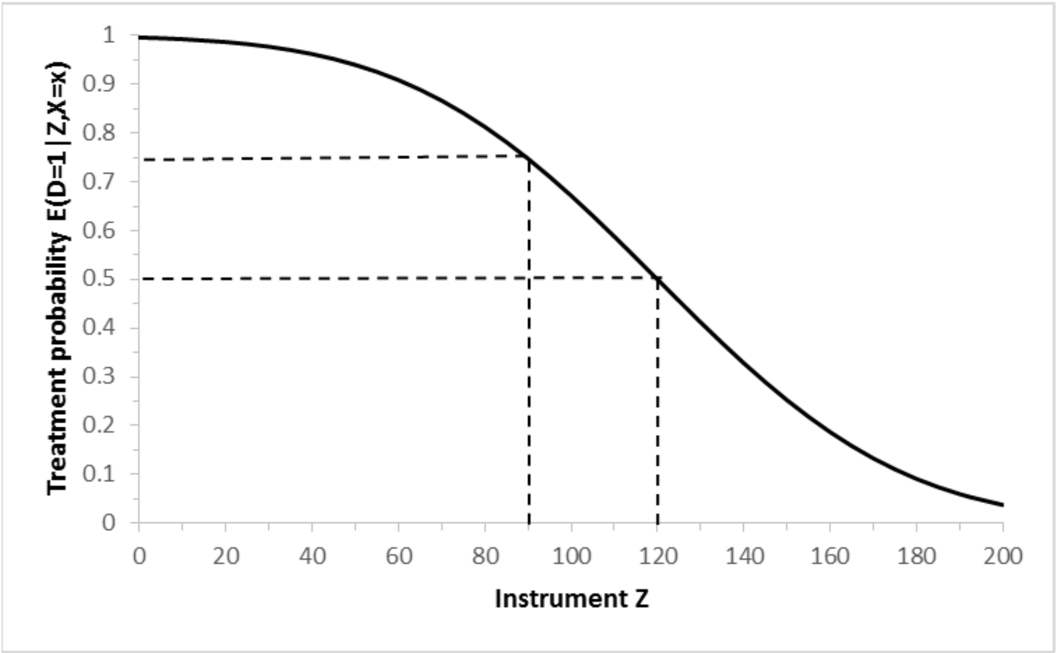
IV with continuous instrument

- If Z is a continuous instrument, can exploit any pair of values z and z' as a binary instrument for pairwise :
  - » Wald$(z, z', x) = \frac{E[Y_i|Z_i=z,X_i=x]-E[Y_i|Z_i=z\prime,X_i=x]}{E[D_i|Z_i=z,X_i=x]-E[D_i|Z_i=z\prime,X_i=x]}$.
- Assuming that move to z to z' shifts individuals into treatment ($E[D|Z = z, X = x] > E[D|Z = z', X = x]$),
  - » $LATE(z, z', x) = E[Y_1 - Y_0|P(z') < \mathrm{U_D} < P(z), X = x]$
- 2SLS: Overall IV effect is variance-weighted average of covariate-specific LATEs (representative for compliers with changes between *all* values of the instrument)

Group of compliers with continuous instrument

- Example: Continuous instrument Z varying between 0 and 200, e.g. Distance to college.
- Decreasing Z from 120 to 90, raises probability of treatment from P(120)=0.5 to P(90)=0.75.
  - This shift individuals with $0.5 < U_D < 0.75$ into treatment

Figure 1: Treatment probability as a function of a continuous instrument



Marginal Treatment Effects (MTE)

- Framework that allows estimating the distribution of treatment effects in the population more fully
- By aggregating over the distribution of marginal treatment effects, one can then recover economically interesting effects (ATE, ATT, ATU, PRTE, etc.) as well as LATE
- First introduced as a parametric normal selection model by Björklund and Moffit (1987) then developed by Heckman, Vytlacil and co-authors in a series of papers, e.g. Heckman and Vytlacil (2005, 2007), Heckman, Urzua and Vytlacil (2006), Carneiro, Heckman and Vytlacil (2011)

Definition of the Marginal Treatment Effect

- The MTE is defined by:

$$MTE(x, u_D) = E(Y_1 - Y_0 | X = x, U_D = u_D)$$
$$= \mu_1(X_i) - \mu_0(X_i) + E(U_1 - U_0 | X = x, U_D = u_D)$$

- That is simply the treatment effect for an individual with observed characteristics x and who is at the u_D-th quantile of the distribution of V (the unobserved "resistance" to treatment)

- For example, and individual with $u_D = .1$ is at the 10th percentile of the distribution of V. If faced with a propensity score of $P(z) = .1$, this individual is just indifferent of taking the treatment.

Identification of the Marginal Treatment Effect

- The MTE for individuals with $X = x$ and $U_D = p$ can be obtained by differentiating the conditional expectation of Y given $X = x$ and $P(Z) = p$ with respect to p:

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

- The derivative of the outcome with respect to the *observed* inducement into treatment (the propensity score) yields the treatment effect for individuals at a given point in the distribution of the *unobserved* resistance to treatment (U_D)

Intuition why MTE is derivative of Y w.r.t. p

- At a given propensity score $p = p_0$, individuals with $U_D < p_0$ are treated, while individuals with $U_D = p_0$ are indifferent
- Increasing p from p_0 by a small amount dp shifts previously indifferent individuals with $U_D = p_0$ into treatment
- They have a marginal treatment effect of MTE($U_D = p_0$)
- The associated increase in Y equals the share of shifted individuals times their treatment effect: *dY=dp* MTE*($U_D = p_0$)
- Dividing the change in Y by the change in p (which is, roughly speaking, what a derivative does) thus gives the MTE: *dY/dp= MTE*($U_D = p_0$)
- Therefore the derivative of the outcome with respect to the propensity score yields the MTE at $U_D = p$.

Estimation of the Marginal Treatment Effect

- Estimating the MTE by the derivative of Y with respect to p is called Local Instrumental Variable (LIV) estimator
- In practice this involves the following steps which can be done more or less parametrically:
  1. Estimate the propensity score $\hat{p}$
     - Probit, Logit, LPM
     - Fully non-parametric (means within all cells of Z)
  2. Model Y as a function of X, $X\hat{p}$ and $\hat{p}$ and estimate it
     - Parametric: A polynomial in $\hat{p}$
     - Semi-parametric: Local linear or quadratic regression in $\hat{p}$, linear in X
  3. Calculate the derivative of $\hat{Y}$ with respect to $\hat{p}$ (at given values of X), and plot it (MTE curve)

Fully Parametric Normal Model

- An alternative to the Local IV Estimator is a fully parametric normal selection model, assuming that U_1, U_0 and V are jointly normally distributed (as in Björklund/Moffitt 1987, Aakvik/Heckman/Vytlacil 2005)
- Outcome and selection equation can then be jointly estimated by Maximum Likelihood, or based on a two-step estimator that plugs selection correction terms computed from the selection equation into the outcome equations
- $\mathrm{E}(U_1 - U_0 | U_D = u_D)$ then has a specific parametric form derived from the joint normal distribution

Estimation of the Marginal Treatment Effect

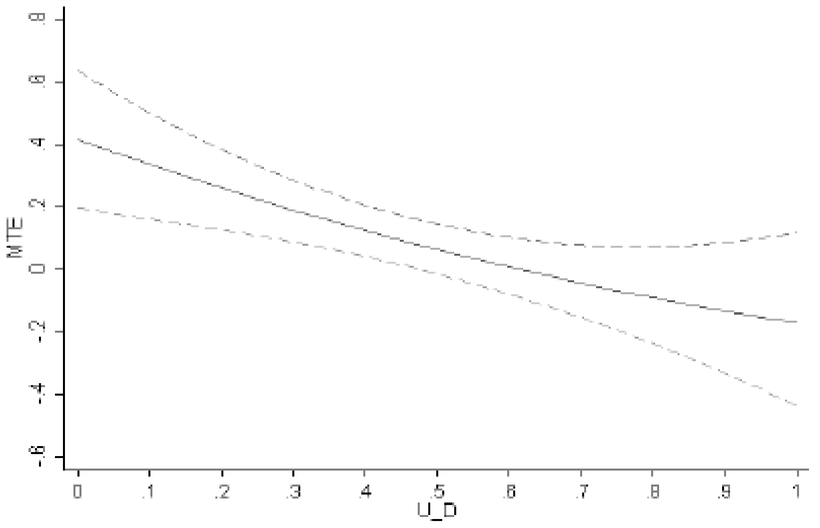
- Practical guidelines for the estimation can be found in the following documents (among others):
  - Appendix B of Heckman, Urzua, Vytlacil (2008, RevEconStat)
  - Heckman, Urzua, Vytlacil 's Practical Guidelines in http://jenni.uchicago.edu/underiv/documentation_2006_03_20.pdf
  - Stata command **`margte`** (Brave/Walstrum 2014, Stata Journal, see http://www.stata-journal.com/sjpdf.html?articlenum=st0331)
  - Stata command **`mtefe`** by Martin Andresen (https://sites.google.com/site/martineckhoffandresen/research)

MTE curve examples

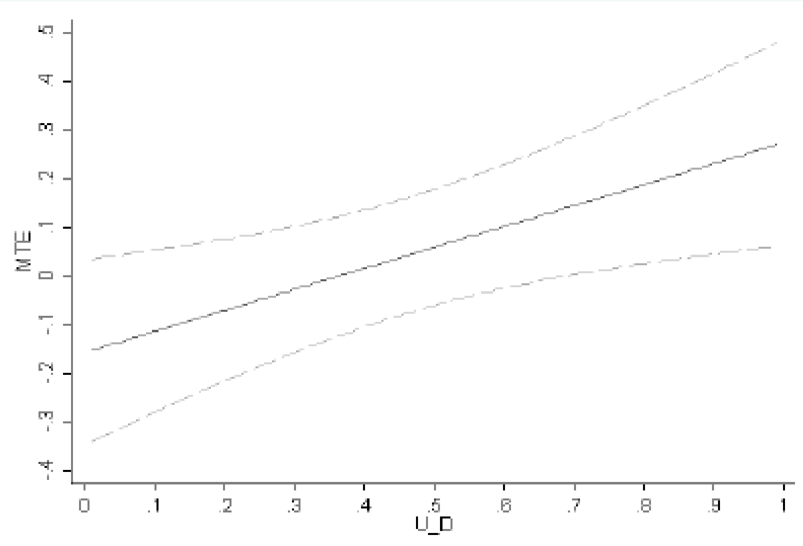
Carneiro et al. (2011, AER)
College attendance

Cornelissen et al. (2018),
Preschool attendance.

Part A: MTE curve for returns to college



Part B: MTE curve for returns to early child care attendance



Notes: Part A depicts the MTE curve of Carneiro, Heckman and Vtylacil (2011, Figure 4) for the wage returns to college estimated by the semi-parametric method (see Appendix B.1). Part B shows the MTE curve of Cornelissen et al. (2016, Figure 4, Part A) for the returns to early child care attendance on school readiness estimated by the parametric polynomial method (see Appendix B.2). In both figures the 90% confidence interval is based on bootstrapped standard errors.

Interpretation of the MTE curve

- Horizontal axis shows quantiles of unobserved resistance to participate in the treatment, vertical axis shows the treatment effect (X are usually held constant at means)
- ATE is an equally weighted average over the MTE curve
- ATT is a weighted average, more heavily weighting individuals to the left (“low resistance = high propensity”)
- ATU is a weighted average, more heavily weighting individuals to the right (“high resistance = low propensity”)
- Downward slope: selection on unobserved gains, ATT>ATE>ATU
- Upward slope: reverse selection on gains, ATU>ATE>ATT

Data requirements

Estimating an MTE curve non-parametrically within all values of $X = x$ requires:

- A continuous instrument Z with sufficient variation conditional on $X = x$ (within all unique combinations of the values of the X's) to generate a propensity score P(Z) with full common support (has support in the full unit interval for both treated and untreated individuals) conditional on $X = x$

This is rarely available. Applications often assume that the shape of the MTE curve does not depend on X and that the outcome is a linear regression in X, estimating the MTE curve semi-parametrically. That eases the requirement to:

- A continuous instrument Z with sufficient variation across $X = x$ to generate a propensity score P(Z) with full common support across $X = x$

Data requirements

If the variation in Z does not induce variation in P(Z) over the full support, one can:

- either estimate the MTE only in the range of support
- or make parametric assumption on the functional form of the MTE curve, and then extrapolate the MTE curve out of the support

Very strong assumptions on the functional form of the MTE curve may not be desirable in many applications.

But they are powerful. For example, a linear MTE curve can be identified with a single dummy variable instrument (Mogstad et al. 2015).

Further References

- Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review*, 106(4): 855-902.
- Heckman, James.1997. “Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations“*The Journal of Human Resources*, 32(3): 441-462
- Oreopoulos, Philip. 2006. "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter." *American Economic Review*, 96(1): 152-175.