Chapter 6: IV beyond LATE

Frontier Topics in Empirical Economics

Lecture notes based on slides by Zibin Huang

College of Business, Shanghai University of Finance and Economics

October 2025

Contents

1 Introduction: The Limits of LATE

In the previous chapter we developed the Local Average Treatment Effect interpretation of IV, following Imbens and Angrist (1994). LATE is elegant, transparent, and tightly linked to a research design: if the instrument is randomly assigned, satisfies exclusion, and induces a monotone change in treatment, then 2SLS recovers the average causal effect for the subpopulation of compliers. Almost every modern IV paper in applied economics is, implicitly or explicitly, an exercise in interpreting an estimate as a LATE.

This is a real achievement, but LATE has two important limitations that the original framework cannot easily overcome (Heckman and Vytlacil, 2007a,b).

First, the standard LATE theorem assumes a single binary instrument and a single binary treatment. In practice we often have multiple instruments, multivalued instruments, or a multivalued treatment such as years of schooling. Once we leave the clean 2×2 world, the IV/2SLS estimand becomes a weighted average of treatment effects, where the weights depend on the first stage, the structure of the instrument, and the behavior of agents. Understanding what 2SLS is actually estimating in these settings requires more work, and in some cases requires us to import economic theory into the identification argument.

Second, LATE is internally valid but not externally valid. The complier subpopulation is defined ex post, by the way agents respond to a particular instrument. If the policy environment changes, the compliers change too. The Vietnam draft compliers of Angrist (1990) are not the same group as the Vietnam volunteers, and they are not the same group as those induced to enlist by a future policy. We can use LATE to evaluate the policy that produced our instrument, but we cannot generally use it to forecast the effect of a new policy on a different group of marginal participants.

This chapter does two things to address these limitations.

Part one generalizes the LATE interpretation within its original design-based framework. We work through three cases in turn: multiple binary instruments, a multivalued (ordered) treatment with a binary instrument, and a multivalued instrument. The first two have clean weighted-average interpretations that follow from imposing monotonicity. The third reveals that monotonicity is not innocuous when the instrument takes more than two values, and forces us to think more carefully about the choice problem agents are solving. We introduce the weak and strong axioms of revealed preference

(WARP and SARP) and walk through the Pinto (2015) Moving to Opportunity (MTO) example to see how economic theory regularizes IV identification when the design alone is not enough.

Part two introduces the Marginal Treatment Effect (MTE) framework of Heckman and Vytlacil (2005, 2007a,b). The MTE is the average treatment effect for agents who are indifferent between taking and not taking treatment, indexed by their observable covariates X and a one-dimensional unobserved resistance U_D. It is a deep structural parameter that does not depend on which instrument we use; LATE, ATE, ATT, ATUT, and the IV and OLS estimands are all weighted averages of the MTE with different weights. We define the MTE and show how to estimate it using the Local Instrumental Variable (LIV) estimand.

We close the chapter with the Kline and Walters (2016) application to Head Start, which illustrates how IV strategy can be interpreted in a more general behavior model perspective and go beyond any single LATE estimate.

The unifying theme of this chapter is that *treatment selection is intrinsically part of any IV problem*. The pure design-based approach hides the selection problem inside the monotonicity assumption. When we let the choice structure come back into the open, IV becomes more powerful, more informative, and more transparently linked to the economic primitives we usually care about.

2 Choice Model Embedded in IV

Before we generalize anything, it is worth recognizing that the LATE framework already contains a choice model. We just used it without naming it.

Recall the LATE theorem from Chapter 5. With binary IV $Z \in \{0, 1\}$ and binary treatment $D \in \{0, 1\}$, agents fall into four behavior types based on the pair $(D(0), D(1))$:

- Always-takers (A): $D(0) = D(1) = 1$.
- Never-takers (N): $D(0) = D(1) = 0$.
- Compliers (C): $D(0) = 0$, $D(1) = 1$.
- Defiers (Def): $D(0) = 1$, $D(1) = 0$.

The monotonicity assumption rules out defiers. Where does this assumption come from? It is not a statistical convenience. It comes from a model of choice. If the instrument Z shifts the cost of treatment, and treatment is a non-Giffen good, then a lower cost can only weakly increase the probability of taking treatment. Formally: if an agent chooses $D = 1$ when the price is high ($Z = 0$), then he or she will also choose $D = 1$ when the price is lower ($Z = 1$). Defiers are ruled out by the law of demand.

This is more than a curiosity. It tells us that whenever we invoke monotonicity, we are implicitly invoking an economic model of how agents respond to incentives. In the 2×2 case the model is so simple it disappears from view. When we move to multivalued instruments and treatments, the underlying choice model has to do more work, and the assumptions we need become richer and more substantive. The slogan of this chapter is exactly that: when statistical tools run out, economics is what you have left.

With this perspective in mind, let us see what happens as we move to more general IV settings.

3 Generalizing LATE: Three Cases

3.1 Recap: Binary Instrument, Binary Treatment

In the 2×2 case there are four observed conditional expectations, one for each combination of Z and D:

$$\begin{aligned}
E[Y \mid Z=1, D=1] &= P(A \mid Z=1, D=1)\, E[Y_1 \mid Z=1, D=1, A] \\
&\quad + P(C \mid Z=1, D=1)\, E[Y_1 \mid Z=1, D=1, C], \\
E[Y \mid Z=1, D=0] &= E[Y_0 \mid Z=1, D=0, N], \\
E[Y \mid Z=0, D=1] &= E[Y_1 \mid Z=0, D=1, A], \\
E[Y \mid Z=0, D=0] &= P(N \mid Z=0, D=0)\, E[Y_0 \mid Z=0, D=0, N] \\
&\quad + P(C \mid Z=0, D=0)\, E[Y_0 \mid Z=0, D=0, C].
\end{aligned}$$

The left hand side is observed; the right hand side mixes potential outcomes that we cannot observe directly. Under monotonicity (no defiers) and the independence of Z, this 4×3 system can be inverted to identify LATE, the average treatment effect for compliers. The tree of behavior types underlying the system is shown in Figure 1.

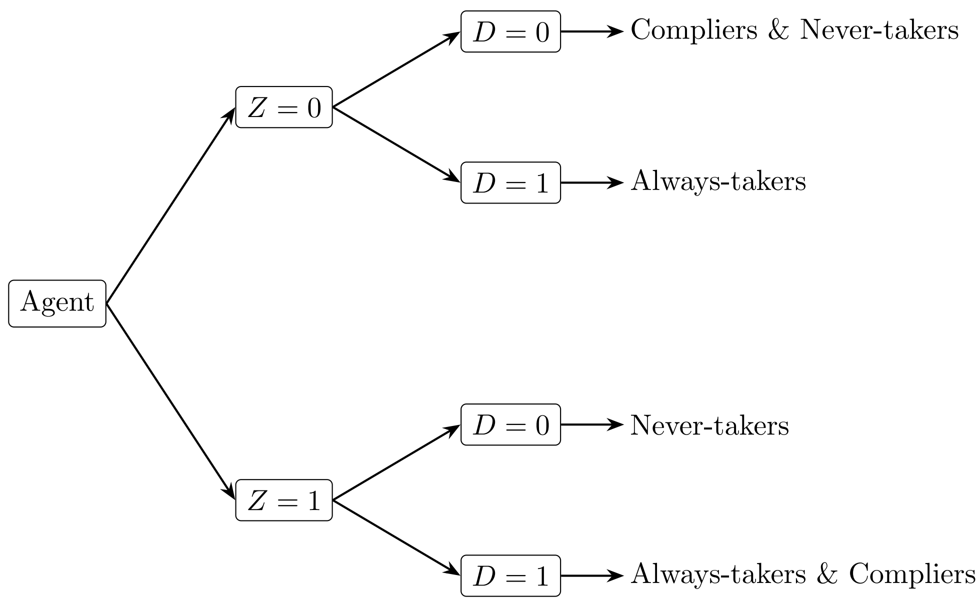

Figure 1: Behavior types observed under each combination of (Z, D) in the binary case, after monotonicity has ruled out defiers.

Three behavior types and four observed expectations: the system is just identifiable enough to recover the LATE for compliers. Once we leave this tight 2×2 world, the number of behavior types grows quickly, and the identification arithmetic becomes more interesting.

3.2 Multiple Binary Instruments

Suppose we have two binary instruments z_1 and z_2, both valid (each independent of potential outcomes and inducing a first stage), and a binary treatment. The natural empirical move is to run 2SLS using both instruments. What does the resulting coefficient represent?

Angrist and Pischke (2009), Chapter 4.5.1, show that if monotonicity holds separately for each instrument, the 2SLS coefficient is a convex combination of the two instrument-specific LATEs:

Proposition 3.1 (2SLS with Multiple Binary Instruments, Chapter 4.5.1 in Angrist and Pischke, 2009)**.** *With two binary instruments z_1, z_2, both satisfying independence and monotonicity, the 2SLS estimand is*

$$\rho_{2SLS} = \psi \, LATE_1 + (1 - \psi) \, LATE_2, \qquad \psi \in (0, 1),$$

where $LATE_k$ is the LATE identified by instrument z_k alone. To express the weight, let the first stage fitted treatment be $\hat{D}_i = \pi_1 z_{1i} + \pi_2 z_{2i}$, where π_1 and π_2 are the first stage coefficients. Then, following Angrist and Pischke (2009, p. 131),

$$\psi = \frac{\pi_1 \operatorname{Cov}(D_i, z_{1i})}{\pi_1 \operatorname{Cov}(D_i, z_{1i}) + \pi_2 \operatorname{Cov}(D_i, z_{2i})} \in (0, 1),$$

so 2SLS gives more weight to the instrument with the stronger first stage.

The intuition is that 2SLS gives more weight to the instrument that does more work in the first stage. A strong instrument with a small LATE will pull the combined estimand toward that small LATE; a weak instrument with a large LATE will contribute less. This is reassuring and a little disquieting at the same time. Reassuring because 2SLS is at least a sensible weighted average; disquieting because the weights are determined by statistical first-stage strength, not by what we care about substantively. If LATE_1 and LATE_2 are similar in magnitude, the combination is informative. If they differ a lot, ρ_{2SLS} can be hard to interpret as a single parameter.

There is also a deeper subtlety. With multiple instruments we have an overidentified system, and we can run the Hansen J-test for overidentifying restrictions (Chapter 5). The test asks whether the two instruments give statistically consistent estimates. Rejecting the test is often interpreted as evidence that at least one instrument is invalid. But with heterogeneous treatment effects, it could equally well mean that the two instruments are valid but identify different LATEs. The J-test cannot distinguish these two stories. In an MTE framework, the second story turns out to be a feature rather than a bug, because the two LATEs correspond to different windows of U_D.

3.3 Multivalued Treatment: The Average Causal Response

Now keep the instrument binary but let the treatment take ordered integer values $s \in \{0, 1, 2, \ldots, \bar{s}\}$. The canonical example is years of schooling. The instrument z might be a compulsory schooling law: $z = 1$ if the law applies, $z = 0$ otherwise. The treatment is the resulting years of schooling, which is naturally multivalued.

Angrist and Pischke (2009) state the relevant generalization, the Average Causal Response (ACR). The assumptions parallel the binary LATE assumptions:

Assumption 3.1 (ACR)**.** Assume that Y_{si} is the potential outcome of Y if treatment is s, and s_{zi} is the potential outcome of s if the instrument is z.

1. **Independence:** $\{Y_{0i}, Y_{1i}, \ldots, Y_{\bar{s},i}; s_{0i}, s_{1i}\} \perp z_i$. The instrument is as good as randomly assigned with respect to all potential outcomes and all potential treatments.
2. **First stage:** $E[s_{1i} - s_{0i}] \neq 0$. The instrument shifts the average level of treatment.
3. **Monotonicity:** $s_{1i} - s_{0i} \geq 0$ for all i (or weakly the reverse for all i). The instrument moves every agent's treatment in the same direction.

The monotonicity assumption now has bite: it requires us to have an ordered list of treatment values, so that the phrase "the instrument moves treatment in the same direction" is meaningful. For schooling this is natural; for unordered categorical treatments it is not.

Under these assumptions, the IV estimand has a clean interpretation.

Theorem 3.1 (Average Causal Response, Theorem 4.5.3 in Angrist and Pischke, 2009)**.** *Under independence, first stage, and monotonicity, the IV/Wald estimator with a binary instrument and ordered multivalued treatment satisfies*

$$\frac{E[Y_i \mid z_i = 1] - E[Y_i \mid z_i = 0]}{E[s_i \mid z_i = 1] - E[s_i \mid z_i = 0]} = \sum_{s=1}^{\bar{s}} \omega_s \, E[Y_{s,i} - Y_{s-1,i} \mid s_{1i} \geq s > s_{0i}],$$

where the weights are

$$\omega_s = \frac{P[s_{1i} \geq s > s_{0i}]}{\sum_{j=1}^{\bar{s}} P[s_{1i} \geq j > s_{0i}]}.$$

Unpack this slowly. The quantity $Y_{s,i} - Y_{s-1,i}$ is the unit response, or stepwise treatment effect, of moving from treatment level $s-1$ to s for individual i. For each step, we average over the compliers whose treatment span crosses that step: agents for whom s_{0i} is below s and s_{1i} is at or above s. The weight ω_s is the share of all step-crossings that occur at step s. By construction the weights sum to one, so the estimand is a proper convex combination of stepwise effects.

This is the ACR: a weighted average of the unit causal responses, over the steps and over the compliers who cross each step. For schooling, an IV estimate using a compulsory schooling law identifies an average return per year of schooling, but only over the marginal years of schooling that the law actually shifts, and only for the compliers who shift on those margins. It is not the population average return per year, and it is not the return for any one fixed margin (say, the return to a four year college degree). What it is, and where it lies in the distribution of returns, depends on which margins the instrument moves.

If we have multiple binary instruments and multivalued treatment, the interpretation combines this case with the previous one: a weighted average across instruments, each of which contributes its own ACR. The weighting scheme becomes baroque quickly, and the practical advice is to be very explicit about what compliers and what margins your instrument is moving.

3.4 Multivalued Instrument and the Monotonicity Trap

Now consider the opposite case: binary treatment but multivalued instrument $z \in \{0, 1, 2, \ldots\}$. An example is the MTO experiment of Pinto (2015), where three voucher policies are randomly assigned. We will return to MTO in detail below. First, let us see why this case is genuinely harder.

A natural first instinct is to decompose the multivalued instrument into binary dummies and run 2SLS. With $z \in \{0, 1, 2\}$ we could define

$$z_1 = \mathbf{1}[z = 1], \qquad z_2 = \mathbf{1}[z = 2],$$

and use both as instruments. Each dummy looks binary, so we are tempted to treat the resulting estimand as a weighted average of two LATEs, as in the multiple-IV case above. Unfortunately monotonicity can fail for the dummies even when it holds for z.

To see this, consider the group with $z_1 = 0$. These agents have either $z = 0$ or $z = 2$. Suppose the original instrument z is monotonic in the natural ordering, so that $D(z = 0) \leq D(z = 1) \leq D(z = 2)$. When z_1 flips from 0 to 1, agents starting from $z = 0$ move *up* to $z = 1$ (treatment weakly increases), but agents starting from $z = 2$ move *down* to $z = 1$ (treatment weakly decreases). The same change in z_1 pushes different agents in opposite directions in D. Monotonicity for z_1 is violated.

This is what we will call the *monotonicity trap*. It is not a small technical issue. It means that the IV estimate using these naive dummies is a weighted combination of effects that may have different signs, and the weights can even be negative. We cannot interpret the resulting estimand as any single average treatment effect for a clean subpopulation.

But is there a different dummy decomposition that could avoid the trap? The answer is yes: instead of indicator dummies, use cumulative dummies $\tilde{z}_1 = \mathbf{1}[z \geq 1]$ and $\tilde{z}_2 = \mathbf{1}[z \geq 2]$. With the natural ordering and monotonicity of z, each cumulative dummy is monotone in D, and the 2SLS estimand is a non-negative convex combination of the LATEs for the marginal step shifts. We leave the verification as a homework problem; it generalizes immediately to instruments with more than three values.

Even with cumulative dummies, however, monotonicity is no longer a free assumption. In the binary case it is essentially harmless: it says the instrument has the same sign for everyone. First, in the multivalued case, it must hold for each step of the instrument, which leads to a set of much stronger assumptions. Second, in many cases instrument values do not have a natural order, leading to difficulty in defining the steps. Third, even if we could decompose the IV successfully, the weighted average of these LATEs is hard to interpret, because for compliers responding to $z > 1$, some are also responding to $z > 2$, which leads to overlap of the complier groups for each sub-LATE before taking the average. As a result, we have to step into the agent's optimization problem to illustrate what we are identifying here.

3.5 Revealed Preference and the Pinto MTO Example

3.5.1 Axioms of Revealed Preference

The natural framework for thinking about how budget shifts (from the instrument) change choices (the treatment) is revealed preference. Two axioms from microeconomics, the Weak Axiom of Revealed Preference (WARP) and the Strong Axiom of Revealed Preference (SARP), turn out to be exactly the tools we need.

Definition 3.1 (WARP, Definition 2.F.1 in Mas-Colell, Whinston, and Green, 1995)**.** The demand function $x(p, w)$ satisfies the Weak Axiom of Revealed Preference if, for any two price-wealth situations (p, w) and (p', w'):

$$p \cdot x(p', w') \leq w \text{ and } x(p', w') \neq x(p, w) \implies p' \cdot x(p, w) > w'.$$

In words: if some bundle is feasible at (p, w) but not chosen, and the agent chose a different bundle there, then that chosen bundle must not be affordable in the new situation. Equivalently: $x_A \succsim_R x_B$.

Definition 3.2 (SARP, Definition 3.J.1 in Mas-Colell, Whinston, and Green, 1995)**.**

The demand function $x(p, w)$ satisfies the Strong Axiom of Revealed Preference if, for any sequence of price-wealth situations $(p^1, w^1), \ldots, (p^N, w^N)$ with $x(p^{n+1}, w^{n+1}) \neq x(p^n, w^n)$ for all $n \leq N-1$, we have $p^N \cdot x(p^1, w^1) > w^N$ whenever $p^n \cdot x(p^{n+1}, w^{n+1}) \leq w^n$ for all $n \leq N-1$.

SARP is WARP plus transitivity. If x_N is revealed preferred to x_{N-1}, which is revealed preferred to x_{N-2}, which is revealed preferred to ..., which is revealed preferred to x_1, then x_N is revealed preferred to x_1. WARP rules out direct cycles between two bundles; SARP rules out cycles of any length.

For the IV identification problem, the SARP machinery lets us carry preferences across different policy environments (different instrument values). If we know an agent chose t over t' at one policy setting (with a corresponding budget), and the budget for t weakly expands and the budget for t' weakly shrinks under a new policy, then SARP says the agent will still choose t over t' at the new budget set. This is exactly the kind of cross-policy restriction we need to rule out implausible behavior types and reduce the number of parameters we have to identify.

3.5.2 The MTO Setting

We illustrate this with the Moving to Opportunity experiment, analyzed by Pinto (2015). MTO randomized vouchers among low income families in five US cities, with the explicit goal of encouraging relocation to lower poverty neighborhoods. There are three random assignments (three values of the instrument):

- Control (z_1): no voucher.
- Experimental (z_2): voucher usable only for housing in a low poverty neighborhood.
- Section 8 (z_3): voucher usable for any housing.

And three possible relocation choices (three values of the treatment):

- $t = 1$: do not relocate.
- $t = 2$: relocate to a low poverty neighborhood.
- $t = 3$: relocate to a high poverty neighborhood.

The realized compliance pattern in MTO is shown in Figure 2, reproducing Figure 1 of Pinto (2015). Most families in the control group did not move; in the experimental group, voucher compliers all moved to low poverty neighborhoods (by construction); and in the Section 8 group, most voucher compliers moved to high poverty neighborhoods. Voucher non-compliers in the experimental and Section 8 groups behaved similarly to the control group, mostly not moving.

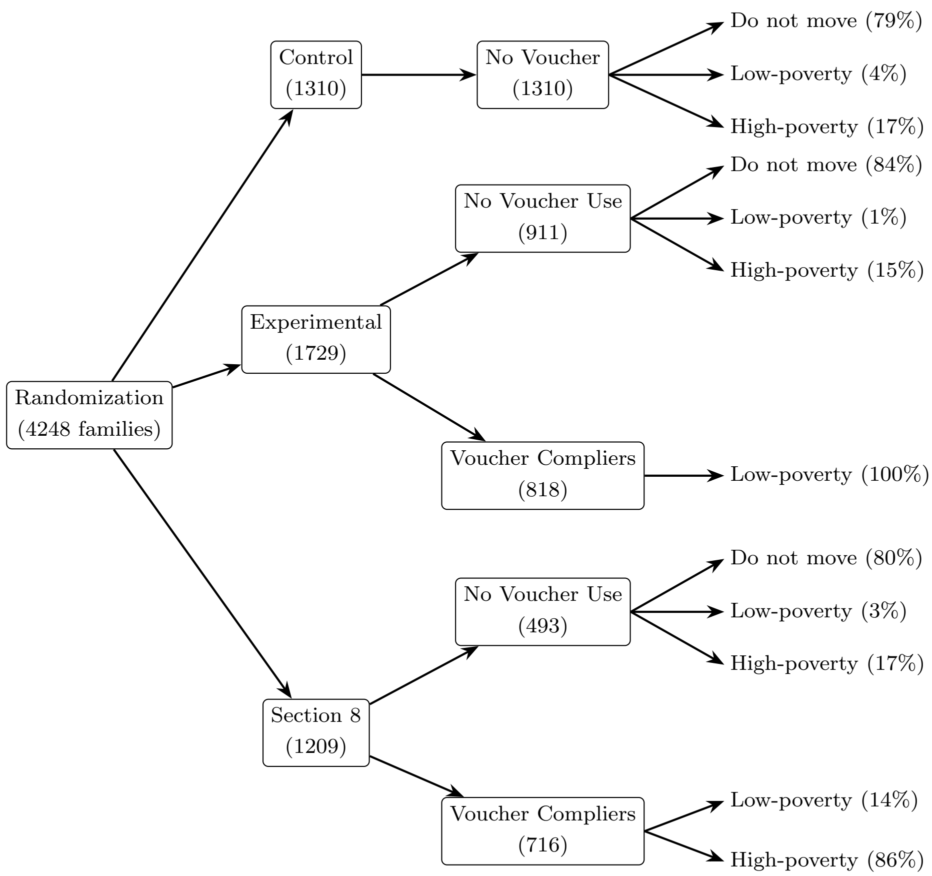

Figure 2: Neighborhood relocation by voucher assignment and compliance in MTO. Adapted from Figure 1 of Pinto (2015).

3.5.3 Why the Type Space Explodes

With three voucher values and three moving choices, an agent's behavior type is described by the triple

$$S_\omega = [C_\omega(z_1), C_\omega(z_2), C_\omega(z_3)],$$

the choice the family would make under each voucher. There are $3^3 = 27$ possible types. We observe nine conditional expectations $E[Y \mid z, t]$ for $z \in \{z_1, z_2, z_3\}$ and $t \in \{1, 2, 3\}$. A system of 9 equations with 27 behavior types cannot identify the underlying causal parameters without further structure. In general, with n choices and m instrument values, the number of behavior types is n^m and the number of observed expectations is $n \times m$. The behavior types grow exponentially in m; the observations grow linearly. There is no hope of identification without ruling out behavior types.

Luckily, the MTO design itself, together with rational choice, lets us rule out most of them. Let $u_\omega(k,t)$ be the family's utility over consumption k and moving choice $t \in \{1,2,3\}$, and let $W_\omega(z,t)$ be the budget set the family faces if it chooses t under voucher z. The three policies translate into the following budget set restrictions.

Assumption 3.2 (A-1 and A-2 in Pinto, 2015)**.** For every family ω:

$$W_\omega(z_1,2) \subsetneq W_\omega(z_2,2) = W_\omega(z_3,2), \tag{1}$$

$$W_\omega(z_1,3) = W_\omega(z_2,3) \subsetneq W_\omega(z_3,3), \tag{2}$$

$$W_\omega(z_1,1) = W_\omega(z_1,2) = W_\omega(z_1,3) = W_\omega(z_2,1) = W_\omega(z_2,3) = W_\omega(z_3,1). \tag{3}$$

In words: (1) under the experimental or Section 8 vouchers, moving to a low poverty neighborhood expands the budget set, because the voucher subsidizes it. (2) Under the Section 8 voucher, moving to a high poverty neighborhood also expands the budget, but the experimental voucher does not subsidize this choice. (3) For any other combination (not moving, or moving to a place the voucher does not subsidize), the budget set is the same as under no voucher.

Combining these with SARP and the standard assumption of rational preferences, Pinto (2015) derives the following choice restrictions.

Lemma 3.1 (Lemma L-1 in Pinto, 2015)**.** *If preferences are rational, then under Assumptions A-1 and A-2:*

1. $C_\omega(z_1) = 2 \Rightarrow C_\omega(z_2) = 2,\ C_\omega(z_3) \neq 1,$
2. $C_\omega(z_1) = 3 \Rightarrow C_\omega(z_2) \neq 1,\ C_\omega(z_3) \neq 1,$
3. $C_\omega(z_2) = 1 \Rightarrow C_\omega(z_1) = 1,\ C_\omega(z_3) \neq 2,$
4. $C_\omega(z_2) = 3 \Rightarrow C_\omega(z_1) = 3,\ C_\omega(z_3) = 3,$
5. $C_\omega(z_3) = 1 \Rightarrow C_\omega(z_1) = 1,\ C_\omega(z_2) = 1,$
6. $C_\omega(z_3) = 2 \Rightarrow C_\omega(z_2) = 2.$

The proof, which we omit, applies SARP repeatedly. The flavor is captured by case 1. If a family chooses to move to a low poverty neighborhood under the control voucher z_1, when no moving option is subsidized, then under the experimental voucher z_2 the budget for moving to a low poverty neighborhood weakly expands while the budgets for not moving and for moving to a high poverty neighborhood do not change. SARP says the family should still prefer moving to a low poverty neighborhood, so $C_\omega(z_2) = 2$. Under z_3, the budgets for both kinds of moves expand, so the family will not choose not to move; but the family might now prefer moving to a high poverty neighborhood, so we cannot rule out $C_\omega(z_3) = 3$. We can only conclude $C_\omega(z_3) \neq 1$.

To go further, Pinto (2015) adds the standard assumption that neighborhoods are normal goods.

Assumption 3.3 (A-3 in Pinto, 2015)**.** For every family ω, and for $z, z' \in \{z_1, z_2, z_3\}$: if $C_\omega(z) = t$ and $W_\omega(z,t) \subsetneq W_\omega(z',t)$, then $C_\omega(z') = t$.

If the family chose t at a smaller budget set and the budget for t strictly expands under a different voucher, the family still chooses t. This rules out cases like $C_\omega(z_1) = 2, C_\omega(z_2) = 2, C_\omega(z_3) = 3$: the family chose moving to low poverty under both z_1 and z_2; under z_3 the budget for moving to low poverty has weakly expanded, so the family should still choose it.

Combining the rules from Lemma L-1 and Assumption A-3, the 27 a priori behavior types collapse to just 7 admissible types. With 9 observed expectations and 7 unknowns, the identification problem becomes manageable.

Lesson

We could not have collapsed 27 to 7 with statistical assumptions alone. Monotonicity in the naive "dummy decomposition" sense was not enough. What did the work was a sequence of economic assumptions: a budget set description of each policy, SARP applied to the underlying preferences, and the normal good restriction. When statistics runs out, microeconomics steps in. Do not throw away your knowledge from the first year Ph.D. microeconomics course.

4 Marginal Treatment Effect (MTE) Framework

4.1 Why MTE? External Validity

We have so far been generalizing LATE inside its original design-based framework. The other limitation we listed at the start of the chapter is external validity. LATE is the average treatment effect for compliers, where compliers are defined by the way agents respond to a specific instrument. If we change the policy, the complier subpopulation changes too. Vietnam draft compliers and Vietnam volunteers are different groups; a hypothetical future draft lottery would identify yet another complier subpopulation.

The problem is that LATE is grouped by ex post behavior, not by ex ante characteristics. We would much rather have a treatment effect indexed by something that the agent brings to the policy, like ability, family background, or unobserved taste, because such ex ante categories are stable across policy environments. The MTE framework, developed by Heckman and Vytlacil (2005, 2007a,b), delivers exactly this. The trick is to write down an explicit model of how the agent decides to take treatment, and to use that model to translate ex post complier categories into ex ante categories indexed by an unobserved resistance to treatment.

4.2 The Latent Index Choice Model

We work with a binary treatment $D \in \{0, 1\}$ and a pair of potential outcomes

$$Y_1 = \mu_1(X, U_1), \tag{4}$$

$$Y_0 = \mu_0(X, U_0), \tag{5}$$

where X is a set of observed covariates and U_0, U_1 are unobserved factors that affect the outcomes in the two treatment states. The treatment choice is governed by a latent index model:

$$D^* = \mu_D(Z) - V, \qquad D = \mathbf{1}[D^* \geq 0]. \tag{6}$$

Z is a vector of instruments, including any elements of X, plus at least one variable that affects the choice but not the outcomes directly. V is the agent's unobserved resistance to treatment. The econometrician observes (X, Z, D, Y) but not (U_0, U_1, V). The three unobservables can be arbitrarily correlated with each other, so selection into treatment is allowed.

Assumption 4.1 (MTE Model)**.**

(A-1) $(U_0, U_1, V) \perp Z \mid X$. The instrument is independent of all unobservables conditional on X.

(A-2) $\mu_D(Z)$ is nondegenerate conditional on X. There is at least one element of Z that varies and is not contained in X, so the instrument actually moves the choice probability.

(A-3) The distribution of V is continuous.

(A-4) $E[|Y_1|]$ and $E[|Y_0|]$ are finite.

(A-5) $0 < \Pr(D = 1 \mid X) < 1$. Treatment status is not deterministically tied to X; both treated and untreated agents exist at every value of X.

(A-1) and (A-2) are the existence of a valid instrument, written in terms of the latent index. (A-3) gives us a continuous propensity score. (A-4) is a regularity condition for taking expectations. (A-5) is the overlap (common support) condition.

4.3 A Roy-Style Example

The latent index model is more concrete than it first appears. Consider the canonical Roy sorting model with two sectors: agriculture ($D = 0$) and a modern sector ($D = 1$). Let Y be a labor market payoff and let Z_1 be an observed relative cost of working in the modern sector. Assume additive separability in the outcome equations:

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

The agent's net utility from choosing the modern sector is the payoff differential minus the relative cost, minus an unobserved utility cost V_C:

$$D^* = \mu_1(X) + U_1 - [\mu_0(X) + U_0] - Z_1 - V_C, \qquad D = \mathbf{1}[D^* \geq 0].$$

Rearranging,

$$D^* = \underbrace{\mu_1(X) - \mu_0(X) - Z_1}_{\mu_D(Z)} + \underbrace{[U_1 - U_0 - V_C]}_{-V},$$

which matches the latent index form (6) with $V = -[U_1 - U_0 - V_C]$. Agents with high $\mu_1 - \mu_0$ (high observable return to the modern sector) or with high $U_1 - U_0$ (high unobservable return) sort into modern. This is positive sorting on returns. The instrument Z_1, an observed cost shifter, only enters $\mu_D(Z)$ and not the outcomes, so it satisfies the exclusion restriction.

4.4 Propensity Score and the Threshold U_D

The next step is to rewrite the choice equation in terms of the propensity score. Define the propensity score

$$P(Z) \equiv \Pr(D = 1 \mid Z, X) = \Pr(\mu_D(Z) > V \mid Z, X) = F_{V|X}(\mu_D(Z)),$$

where $F_{V|X}$ is the conditional CDF of V given X. The propensity score is just the probability that an agent with instruments Z takes treatment.

Apply the same CDF to the unobserved resistance:

$$U_D \equiv F_{V|X}(V) \sim \text{Uniform}[0, 1].$$

U_D is a uniform random variable that records the agent's quantile in the conditional distribution of V. It is also the threshold propensity score the agent has to pass to get treated when he or she draws resistance V. An agent with $U_D = 0.2$ has lower than average resistance to treatment; an agent with $U_D = 0.9$ has high resistance. Because $F_{V|X}$ is strictly monotonic (under continuity), V and U_D carry the same information, but U_D is normalized to lie in the unit interval.

Now compare the propensity score and the threshold:

$$\mu_D(Z) > V \iff \underbrace{F_{V|X}(\mu_D(Z))}_{P(Z)} > \underbrace{F_{V|X}(V)}_{U_D}.$$

The agent takes treatment if and only if the propensity score from Z exceeds the agent's quantile of resistance. We can therefore characterize each agent by the pair (X, U_D) rather than (X, V). The unobservable is now bounded in $[0, 1]$ and uniform by construction. This re-parameterization is the technical move that makes the rest of the framework work.

4.5 Vytlacil Equivalence

A central result is due to Vytlacil (2002): under assumptions (A-1) to (A-5), the additively separable latent index model is observationally equivalent to the LATE model of Imbens and Angrist (1994). In particular, the latent index model automatically implies monotonicity.

The intuition is that the choice equation $D^* = \mu_D(Z) - V$ is additively separable in Z and V. Given any two instrument values z and z',

$$D^*(z) - D^*(z') = \mu_D(z) - \mu_D(z'),$$

which has the same sign for every agent regardless of V. If $\mu_D(z) > \mu_D(z')$, every agent's latent utility for treatment goes up when we move from z' to z; nobody can go in the opposite direction. That is exactly monotonicity.

The equivalence works in both directions: the latent index model with (A-1) to (A-5) implies the LATE assumptions, and conversely the LATE assumptions can be represented by a latent index model. So MTE is not an alternative framework that replaces LATE; it is a richer parameterization of the same identification setup. The benefit is that the latent index gives us a continuous index U_D along which we can disaggregate the treatment effect.

4.6 Defining the Marginal Treatment Effect

Let $\Delta = Y_1 - Y_0$ denote the individual treatment effect. The conditional ATE is the population mean

$$\Delta^{\text{ATE}}(x) \equiv E[\Delta \mid X = x].$$

The Marginal Treatment Effect localizes this by conditioning on the unobservable as well.

Definition 4.1 (MTE, Heckman and Vytlacil, 2005)**.** The Marginal Treatment Effect at observable x and unobserved resistance u_D is

$$\Delta^{\text{MTE}}(x, u_D) \equiv E[Y_1 - Y_0 \mid X = x, U_D = u_D].$$

This is the average treatment effect for agents who share covariates x and quantile u_D

of unobserved resistance. Equivalently, it is the average effect for agents who would be exactly indifferent between treatment and no treatment if they were assigned an instrument value z with $P(z) = u_D$. They are "marginal" in the textbook sense: at the threshold.

A few features make MTE attractive.

First, the MTE is indexed by ex ante characteristics. x describes the observable type and u_D describes the unobserved appetite for treatment. Neither depends on the particular instrument used in any one study. If we change the policy, the same agent has the same (x, u_D), and the same MTE value.

Second, the MTE traces out treatment effect heterogeneity along u_D. In a positive-selection Roy model, agents with low u_D are those most eager to be treated; they tend to have high gains. Agents with high u_D are reluctant to be treated and tend to have low or negative gains. The MTE curve as a function of u_D is typically decreasing in such settings, and a flat MTE corresponds to homogeneous treatment effects.

Third, and most importantly for empirical work, the MTE is a deep structural object that does not depend on the instrument. LATE, ATE, ATT, ATUT, and even the OLS estimand can all be written as weighted averages of $\Delta^{\text{MTE}}(x, u_D)$ with different weights. If we know the MTE, we know all of these.

4.7 MTE as a Unified Framework

The general representation is

$$\text{TE}(j) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_j(x, u_D)\, du_D,$$

where j indexes the treatment parameter (ATE, ATT, LATE, etc.) and ω_j is the parameter-specific weight. Four leading cases:

ATE. $\omega_{\text{ATE}}(x, u_D) = 1$. The ATE simply integrates the MTE over the full range of U_D.

ATT (Treatment on the Treated). The weight ω_{ATT} puts more mass on agents with low u_D, who are over-represented among the treated under positive selection. The ATT is typically larger than the ATE in a positive sorting model.

ATUT (Treatment on the Untreated). The weight ω_{ATUT} puts more mass on agents with high u_D, who are over-represented among the untreated. Under positive selection, ATUT is below the ATE.

LATE. For an instrument that shifts the propensity score from u'_D to $u_D > u'_D$, the LATE is the average of the MTE over the induced window:

$$\begin{aligned}\text{LATE} &= E[Y_1 - Y_0 \mid X = x, D(z) = 1, D(z') = 0] \\ &= E[Y_1 - Y_0 \mid X = x, u'_D < U_D \leq u_D] \\ &= \int_{u'_D}^{u_D} \Delta^{\text{MTE}}(x, u)\, du. \end{aligned} \tag{7}$$

A LATE is the integral of the MTE over the window of U_D that the instrument actually moves. Different instruments move different windows, which is exactly why different instruments give different LATEs.

Figure 3 shows a stylized picture, recreating Figure 1 of Heckman and Vytlacil (2005), of the MTE together with the ATE, the ATT weight, and the ATUT weight as functions of u_D in a positive-selection Roy model. The MTE declines in u_D (high resistance agents have low gains), the ATT weight is concentrated at low u_D, the ATUT weight is concentrated at high u_D, and the ATE is the unweighted area under the MTE curve.

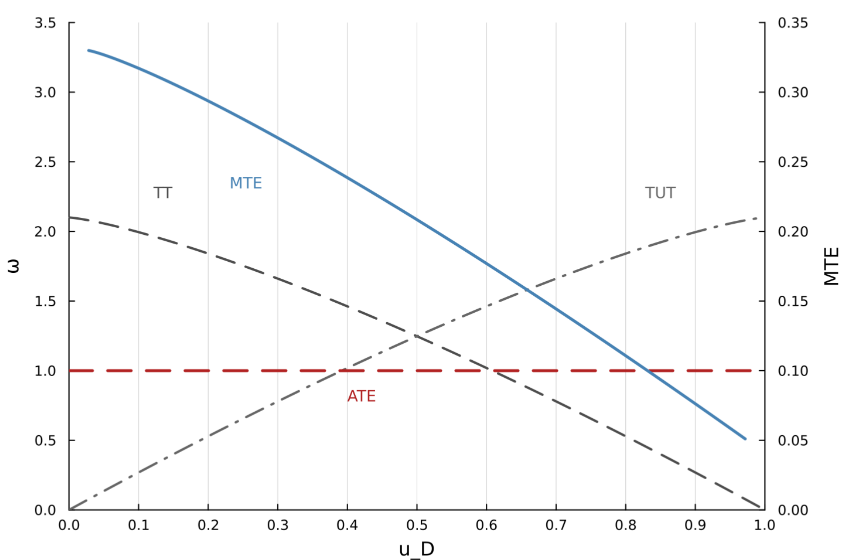

Figure 3: Stylized MTE curve with ATE reference line and ATT/ATUT weights in a positive-selection Roy model. Recreated from Figure 1 of Heckman and Vytlacil (2005).

The picture makes clear that the various treatment parameters are not competitors but views of the same object. They differ only in which part of the MTE curve they emphasize, which is determined by the weight function. The MTE itself is policy invariant: the underlying schedule does not move as the instrument changes.

4.8 Estimation via the Local Instrumental Variable

How do we estimate the MTE from data? The key insight, again from Heckman and Vytlacil (2005), is that the MTE can be recovered as the derivative of a conditional mean function. Suppress the conditioning on X for clarity and write the propensity score as $P(Z)$. Define the **Local Instrumental Variable (LIV)** as

$$\Delta^{\text{LIV}}(p) \equiv \frac{\partial E[Y \mid P(Z) = p]}{\partial p}.$$

LIV is the local response of the outcome to a small change in the propensity score.

Theorem 4.1 (MTE-LIV Equivalence, Heckman and Vytlacil, 2005)**.** *Under assumptions (A-1) to (A-5),*

$$\Delta^{MTE}(p) = \Delta^{LIV}(p) = \frac{\partial E[Y \mid P(Z) = p]}{\partial p}.$$

The intuition is clean. A small policy nudge that raises the propensity score from p to $p + dp$ only changes the treatment status of marginal agents whose threshold is exactly $u_D = p$. Among those marginal agents, the change in Y is precisely their treatment effect, which is $\Delta^{\text{MTE}}(p)$. Aggregating, the derivative of $E[Y \mid P]$ at p delivers the MTE at p.

In practice the estimation runs in three steps.

Step 1. Estimate a treatment choice model. A probit or logit regression of D on Z delivers $\hat{P}(Z)$, the estimated propensity score.

Step 2. Estimate $E[Y \mid X, P(Z)]$ as a function of $\hat{P}$. This can be done flexibly: a series expansion in $\hat{P}$, a kernel smoother, a local polynomial. With a linear specification one obtains a constant MTE; with a flexible specification one obtains a nontrivial MTE curve.

Step 3. Take the derivative with respect to $\hat{P}$. If the second stage uses a polynomial of order K in $\hat{P}$, the MTE is given by the derivative polynomial. With nonparametric smoothing, the derivative is read off the slope of the conditional mean function.

Alternatively, the entire model can be estimated parametrically. Kline and Walters (2016) write a fully parametric MTE model and estimate the joint distribution of (U_0, U_1, V) via maximum likelihood. The parametric approach is more efficient when the assumptions hold and is the natural choice when the goal is policy simulation.

4.9 Implementation: The `mtefe` Stata Package

The MTE framework can be implemented in Stata using the user written package `mtefe` by Martin Andresen. `mtefe` fits a parametric or semiparametric MTE model and returns a battery of treatment effect estimates: ATE, ATT, ATUT, LATE, and Policy Relevant Treatment Effects (PRTE) for hypothetical policies. The package also produces tests for observable and unobserved heterogeneity. Typical output for a wage equation with experience controls and district fixed effects, with a district-level cost shifter as the instrument, includes effect estimates of the form

- ATE around 0.33,
- ATT around 0.54 (treated agents have above average gains, consistent with positive sorting),
- ATUT around 0.12 (untreated agents have lower gains),
- LATE close to the ATE in this stylized example.

Figure 4 illustrates the typical `mtefe` output. The left panel shows the common support of the propensity score for treated and untreated subsamples; the right panel shows the estimated MTE as a function of u_D together with a 95% confidence band and the ATE reference line.

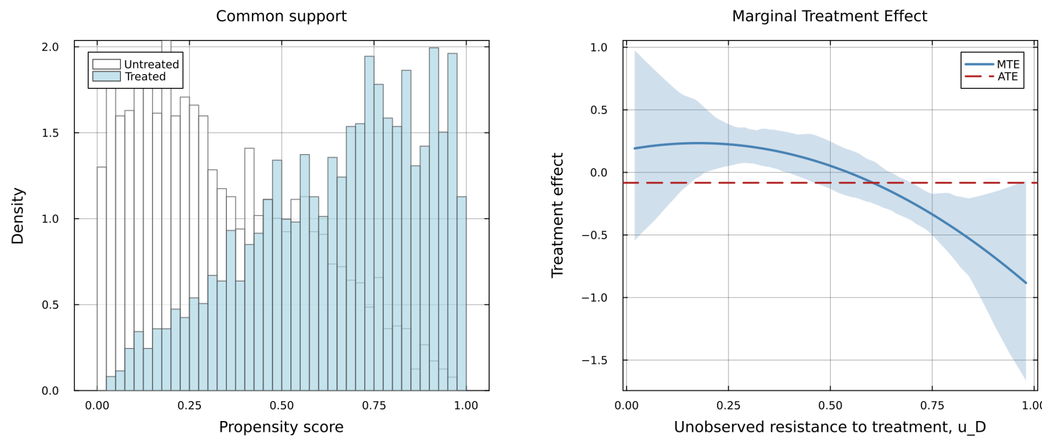

Figure 4: Stylized `mtefe` output. Left: propensity score density by treatment status. Right: estimated MTE curve with 95% confidence band and ATE reference line. The downward sloping MTE indicates positive selection on unobserved gains.

5 Application: Kline and Walters (2016)

Kline and Walters (2016) apply the MTE framework to evaluate the Head Start program, an early childhood education program for low income families in the United States. Their study is an instructive illustration of how the MTE delivers what neither a single LATE nor an observational comparison can deliver on its own.

Head Start enrollment looked very different in two kinds of studies. Observational comparisons of Head Start participants to non-participants showed large positive effects on later outcomes. The Head Start Impact Study (HSIS), a randomized control trial, showed small and often statistically insignificant effects. The natural inference is that observational studies suffered from selection bias and the RCT is closer to the truth. Kline and Walters (2016) argue this inference misses the point.

The reason is that the treatment in Head Start is not really binary. Families have three options: no preschool program at all, Head Start, or a close substitute (typically another subsidized or private preschool). Observational comparisons effectively contrast Head Start with the "no program" option. The RCT contrasts the assigned Head Start option with whatever counterfactual the control group can find, which often includes another preschool program. When close substitutes are available, the LATE identified by the RCT is the effect of Head Start versus the average alternative, not the effect of Head Start versus nothing. The two studies estimate different parameters and the apparent contradiction dissolves.

Kline and Walters (2016) formalize this with a three valued treatment (no program, other program, Head Start) and use the MTO machinery from earlier in this chapter to characterize the admissible behavior types under the HSIS lottery. They estimate a parametric MTE model and report a rich set of causal parameters: ITT, LATE, and several MTE based objects that are externally valid in the sense that their interpretation does not depend on the specific compliance pattern of the HSIS sample.

The headline policy result is that, once we account for substitution from other preschool programs to Head Start, the program is far more cost effective than the small RCT impact would suggest. From the government's perspective, a family who switches from a different subsidized program to Head Start carries with them a savings on the alternative program. This savings, combined with the (modest) marginal benefit of Head Start over the substitute, gives Head Start a positive net social return.

The methodological lesson is the one we have been building toward. A single LATE estimate, no matter how cleanly identified, can be misleading when the treatment effect is heterogeneous and policy environments change the composition of compliers. The MTE framework lets us write down the policy question in terms of stable structural parameters, estimate those parameters from one experiment, and use them to forecast the effect of related but distinct policies.

6 Conclusion

LATE is internally valid and elegantly tied to a research design, but it leaves two questions open. What does 2SLS identify when the instrument or the treatment is not binary? And what does a LATE estimated under one policy tell us about a different policy?

The first half of this chapter addressed the first question by working through three cases: multiple binary instruments (a ψ-weighted average of instrument-specific LATEs), multivalued ordered treatment with a binary instrument (the Average Causal Response of stepwise effects), and multivalued instrument (where naive dummy decomposition breaks monotonicity). The multivalued instrument case forced us to import economic theory through WARP and SARP, and the MTO example of Pinto (2015) showed concretely how budget set restrictions plus rational choice cut the behavior type space down from $3^3 = 27$ to a tractable 7.

The second half introduced the MTE framework of Heckman and Vytlacil (2005, 2007a,b). By writing down an explicit latent index choice model and re-parameterizing the unobserved resistance to lie in the unit interval, we obtain a treatment effect $\Delta^{\text{MTE}}(x, u_D)$ that is indexed by ex ante observables and ex ante unobservables, and that does not depend on which instrument we use. ATE, ATT, ATUT, LATE, IV, and OLS estimands are all weighted averages of the MTE. We can estimate the MTE through the Local Instrumental Variable, which is the derivative of $E[Y \mid P(Z)]$ with respect to the propensity score, using either nonparametric smoothing or fully parametric maximum likelihood. Kline and Walters (2016) illustrate the framework with Head Start.

The thread running through both halves is the same: treatment selection is intrinsic to any IV exercise, and ignoring it means leaving information on the table. Importing economic theory back into IV regularizes identification in hard cases, generalizes interpretation to richer policy settings, and makes the resulting estimates more useful for forecasting future policies.

A Homework Problems

1. **Cumulative-Dummy Decomposition of a Multivalued Instrument.**

   Suppose the instrument z takes values in $\{0, 1, 2\}$ and the treatment D is binary. Assume the underlying instrument is monotonic, so $D(z = 0) \leq D(z = 1) \leq D(z = 2)$ for every agent.

   (a) Define the indicator dummies $z_1 = \mathbf{1}[z = 1]$ and $z_2 = \mathbf{1}[z = 2]$. Show that they do not jointly satisfy monotonicity for D, and identify a behavior pattern that violates monotonicity for z_1.

   (b) Now define cumulative dummies $\tilde{z}_1 = \mathbf{1}[z \geq 1]$ and $\tilde{z}_2 = \mathbf{1}[z \geq 2]$. Show that each $\tilde{z}_k$ is monotonic in D.

   (c) Argue that 2SLS using these cumulative dummies as instruments identifies a non-negative weighted average of the two stepwise LATEs: the LATE for the $z = 0 \to 1$ shift and the LATE for the $z = 1 \to 2$ shift.

   (d) Generalize: describe the cumulative dummy decomposition for an instrument with $K+1$ values.

2. **Choice Restrictions in the Pinto MTO Example.**

   Verify two of the six implications in Lemma L-1 of Pinto (2015), restated in the notes. You may pick any two of the six. For each, state the relevant budget set comparisons under Assumptions A-1 and A-2, and explain how SARP delivers the implication.

3. **LATE as the Integral of the MTE.**

   Consider the latent index model in Section 4 with all of (A-1) to (A-5) satisfied. Let z and z' be two instrument values with $P(z) > P(z')$.

   (a) Show that the LATE identified by the instrument moving from z' to z equals

   $$\text{LATE} = \int_{P(z')}^{P(z)} \Delta^{\text{MTE}}(u)\, du \Big/ [P(z) - P(z')].$$

   (Note the normalization by the width of the propensity-score window.)

   (b) Suppose the MTE is constant at $\Delta^{\text{MTE}}(u) = c$ for all u. Show that every LATE equals c, and that the ATE, ATT, and ATUT also equal c.

Practice Example: Weighted LATE with Two Binary Instruments

This practice example illustrates the multiple binary instruments case from Section 3.2. We simulate a population in which two binary instruments identify two distinct LATEs, and we verify that the 2SLS coefficient is a convex combination of them, weighted by the relative first stage strengths.

Data generating process. Consider $N = 5000$ agents, each with a binary treatment D and a continuous outcome Y. Two binary instruments z_1 and z_2 are independently assigned with $\Pr(z_k = 1) = 0.5$. Each instrument shifts a distinct subgroup of agents from $D = 0$ to $D = 1$. Subgroup A consists of 30% of agents who respond only to z_1 with true treatment effect $\beta_A = 1.0$. Subgroup B consists of 30% of agents who respond only to z_2 with true treatment effect $\beta_B = 3.0$. The remaining 40% are never-takers or always-takers and do not respond to either instrument. The instruments z_1 and z_2 are independent of potential outcomes by construction.

Predicted behavior. The Wald estimator using z_1 alone should recover $\text{LATE}_1 = \beta_A = 1.0$; using z_2 alone, $\text{LATE}_2 = \beta_B = 3.0$. The 2SLS estimator using both should produce

$$\rho_{\text{2SLS}} = \psi \cdot 1.0 + (1 - \psi) \cdot 3.0,$$

with ψ determined by the first stage strengths. Because the two subgroups are the same size, ψ should be close to 0.5 and the 2SLS estimate close to 2.0.

Analysis. The Stata do-file (`src/chapter 6/ch6_iv_beyond_late.do`) performs the following steps:

1. Simulate the population with two distinct complier subgroups.
2. Compute the two Wald estimators using each instrument separately. Confirm $\text{LATE}_1 \approx 1$ and $\text{LATE}_2 \approx 3$.
3. Run 2SLS using both instruments jointly. Verify the coefficient is a convex combination of the two LATEs.
4. Recover the implied weight $\hat{\psi}$ by inverting the convex combination formula, and report it alongside the first stage F statistics.
5. Discuss interpretation: 2SLS gives more weight to the instrument with the stronger

first stage.

Key lessons.

- With heterogeneous treatment effects and multiple instruments, 2SLS does not identify any one LATE: it identifies a convex combination.
- The combination weights are determined by first stage strength, not by what the researcher cares about. This is reassuring and slightly disquieting at the same time.
- In an MTE framework, the two LATEs correspond to different windows of U_D. The fact that they differ is not evidence of an invalid instrument; it is the heterogeneous structure of treatment effects.

The do file is located at `src/chapter 6/ch6_iv_beyond_late.do` and the simulated dataset at `data/chapter 6/ch6_iv_beyond_late.dta`.

References

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*, 80(3): 313–336.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

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

Heckman, James J. and Edward J. Vytlacil. 2007a. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." *Handbook of Econometrics*, 6: 4779–4874.

Heckman, James J. and Edward J. Vytlacil. 2007b. "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." *Handbook of Econometrics*, 6: 4875–5143.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Kline, Patrick and Christopher R. Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *Quarterly Journal of Economics*, 131(4): 1795–1848.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. Oxford University Press.

Pinto, Rodrigo. 2015. "Selection Bias in a Controlled Experiment: The Case of Moving to Opportunity." Unpublished Ph.D. Thesis, University of Chicago, Department of Economics.

Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica*, 70(1): 331–341.