Understanding Instrumental Variables in Models with Essential Heterogeneity

James Heckman, Sergio Urzua and Edward Vytlacil

Econ 312, Spring 2023
This draft, April 17, 2023

Policy adoption problem

- Suppose a policy is proposed for adoption in a country.
- What can we conclude about the likely effectiveness of the policy in countries?
- Build a model of counterfactuals.

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1 \\ Y_0 &= \mu_0(X) + U_{0.} \end{aligned} \tag{1}$$

Consider the basic generalized Roy model

- Two potential outcomes (Y_0, Y_1).
- A choice equation

$$D = \mathbf{1}[\underbrace{\mu_D(Z, V)}_{\text{net utility}} > 0].$$

- Observed outcomes are

$$Y = DY_1 + (1 - D)Y_0$$

- Assume $\mu_D(Z, V) = \mu_D(Z) - V$.

Switching Regression Notation

$$\begin{aligned} Y &= Y_0 + (Y_1 - Y_0)D \qquad (2) \\ &= \mu_0 + (\mu_1 - \mu_0 + U_1 - U_0)D + U_0. \end{aligned}$$

(Quandt, 1958, 1972)

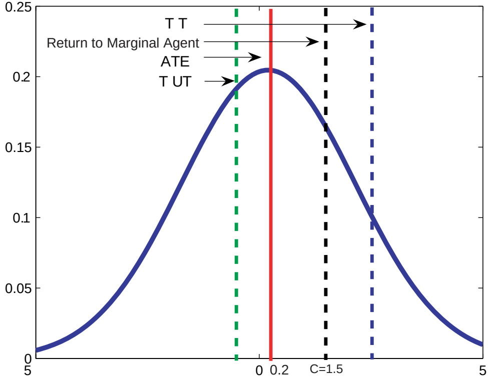
In Conventional Regression Notation

$$Y = \alpha + \beta D + \varepsilon \qquad (3)$$

$\alpha = \mu_0$, $\beta = (Y_1 - Y_0) = \mu_1 - \mu_0 + U_1 - U_0$, $\varepsilon = U_0$.

- β is the "treatment effect."

Figure 1: distribution of gains, a Roy economy



$$\beta = Y_1 - Y_0$$

$$\text{TT} = 2.666,\ \text{TUT} = -0.632$$

$$\text{Return to Marginal Agent} = C = 1.5,\ \text{ATE} = \mu_1 - \mu_0 = \bar{\beta} = 0.2$$

The model

Outcomes	Choice Model
$Y_1 = \mu_1 + U_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \mu_0 + U_0 = \alpha + U_0$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$

General Case

$$(U_1 - U_0) \not\perp\!\!\!\perp D$$
$$\text{ATE} \neq \text{TT} \neq \text{TUT}$$

The model

The Researcher Observes (Y, D, C)

$$Y = \alpha + \beta D + U_0 \text{ where } \beta = Y_1 - Y_0$$

Parameterization

$$\alpha = 0.67 \quad (U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma}) \quad D^* = Y_1 - Y_0 - C$$

$$\bar{\beta} = 0.2 \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \quad C = 1.5$$

- In the case when $U_1 = U_0 = \varepsilon_0$, simple least squares regression of Y on D subject to a **selection bias**.
- This is a form of endogeneity bias considered by the Cowles analysts.
- Upward biased for β if $\mathrm{Cov}(D, \varepsilon) > 0$.

- Three main approaches have been adopted to solve this problem:
  1. Selection models
  2. Instrumental variable models
  3. Matching: assumes that $\varepsilon \perp\!\!\!\perp D \mid X$.
- Matching is just nonparametric least squares and assumes access to rich data which happens to guarantee this condition.

Case I, the traditional case: β is a constant

- If there is an instrument Z, with the property that

$$\mathrm{Cov}(Z, D) \neq 0 \tag{4}$$
$$\mathrm{Cov}(Z, \varepsilon) = 0, \tag{5}$$

then

$$\operatorname{plim} \hat{\beta}_{\mathrm{IV}} = \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, D)} = \beta.$$

- If other instruments exist, each identifies the same β.

Case II, heterogeneous response case: β is a random variable even conditioning on X

Sorting bias or sorting on the gain which is distinct from sorting on the level.

Essential heterogeneity

$$\mathrm{Cov}(\beta, D) \neq 0.$$

Suppose (4), (5) and

$$\mathrm{Cov}(Z, \beta) = 0. \tag{6}$$

- Can we identify the mean of $(Y_1 - Y_0)$ using IV?

- In general we cannot (Heckman and Robb, 1985).
- Let

$$\bar{\beta} = (\mu_1 - \mu_0)$$
$$\beta = \bar{\beta} + \eta$$
$$U_1 - U_0 = \eta$$
$$Y = \alpha + \bar{\beta} D + [\varepsilon + \eta D] .$$

- Need Z to be uncorrelated with $[\varepsilon + \eta D]$ to use IV to identify $\bar{\beta}$.
- This condition will be satisfied if policy adoption is made without knowledge of $\eta \, (= U_1 - U_0)$.
- If decisions about D are made with partial or full knowledge of η, IV does not identify $\bar{\beta}$.

- The IV condition is

$$E\left[\varepsilon + \eta D \mid Z\right] = 0.$$

- $E\left(\varepsilon \mid Z\right) = 0, \quad E\left(\eta \mid Z\right) = 0.$
- Even if $\eta \perp\!\!\!\perp Z$, $\eta \not\perp\!\!\!\perp Z \mid D = 1$.
- $E(\eta D \mid Z) = E(\eta \mid D = 1, Z) \Pr(D = 1 \mid Z)$.
- But $E\left(\eta \mid Z, D = 1\right) \neq 0$, in general, if agents have some information about the gains.

- Draft Lottery example (Heckman, 1997).
- Linear IV does not identify ATE or any standard treatment parameters.

Imbens Angrist conditions (1994)

- Imbens and Angrist (1994) establish that IV can identify an interpretable parameter in the model with essential heterogeneity.
- Their parameter is a discrete approximation to the marginal gain parameter of Björklund and Moffitt (1987).
- This parameter can be interpreted as the marginal gain to outcomes induced from a marginal change in the costs of participating in treatment (Björklund-Moffitt).

Imbens Angrist conditions (1994)

- Imbens and Angrist assume the existence of an instrument Z that takes two or more distinct values.
- Keep conditioning on X implicit.
- Let $D_i(z)$ be the indicator ($= 1$ if adopted; $= 0$ if not)
- It is a random variable for choice when we set $Z = z$.

Imbens Angrist conditions (1994)

(IV-1) (Independence)

$Z \perp\!\!\!\perp \left(Y_1, Y_0, \{D(z)\}_{z \in \mathcal{Z}}\right).$

(IV-2) (Rank)

$\Pr(D = 1 \mid Z)$ *depends on* Z.

- They supplement the standard *IV* assumption with a "monotonicity" assumption.

(IV-3) (Monotonicity or Uniformity)

$D_i(z) \geq D_i(z')$ *or* $D_i(z) \leq D_i(z')$ $i = 1, \ldots, I$.

Imbens Angrist conditions (1994)

- *Uniformity* of responses *across* persons.
- Uniformity is satisfied when, for $z < z'$, $D_i(z) \leq D_i(z')$ for all i, while for $z'' > z'$, $D_i(z'') \leq D_i(z')$ for all i.

Imbens Angrist conditions (1994)

- These conditions imply the LATE parameter.

$$
\begin{aligned}
& E\left(Y \mid Z=z\right)-E\left(Y \mid Z=z'\right) \\
& \quad = E\left(\left(D(z)-D(z')\right)\left(Y_1-Y_0\right)\right) \qquad \text{(Independence)}
\end{aligned}
$$

Imbens Angrist conditions (1994)

- Using iterated expectations,

$$
\begin{aligned}
& E\left(Y \mid Z=z\right)-E\left(Y \mid Z=z'\right) \\
&= \left(\begin{array}{c} E\left(Y_1-Y_0 \mid D\left(z\right)-D\left(z'\right)=1\right) \\ \cdot \Pr\left(D\left(z\right)-D\left(z'\right)=1\right) \end{array}\right) \\
&\quad - \left(\begin{array}{c} E\left(Y_1-Y_0 \mid D\left(z\right)-D\left(z'\right)=-1\right) \\ \cdot \Pr\left(D\left(z\right)-D\left(z'\right)=-1\right) \end{array}\right).
\end{aligned} \tag{7}
$$

- Monotonicity allows us to drop out one term.

Imbens Angrist conditions (1994)

- Suppose, for example, that $\Pr(D(z) - D(z') = -1) = 0$. Thus,

$$
\begin{aligned}
& E\left(Y \mid Z = z\right) - E\left(Y \mid Z = z'\right) \\
& \quad = E\left(Y_1 - Y_0 \mid D(z) - D(z') = 1\right) \Pr\left(D(z) - D(z') = 1\right).
\end{aligned}
$$

$$
\begin{aligned}
LATE &= \frac{E\left(Y \mid Z = z\right) - E\left(Y \mid Z = z'\right)}{\Pr(D = 1 \mid Z = z) - \Pr(D = 1 \mid Z = z')} \\
&= E\left(Y_1 - Y_0 \mid D(z) - D(z') = 1\right) \qquad (8)
\end{aligned}
$$

- The mean gain to those induced to switch from "0" to "1" by a change in Z from z' to z.

Imbens Angrist conditions (1994)

- Observe LATE = ATE if

$$\Pr\left(D=1 \mid Z=z\right)=1 \quad \text{while} \quad \Pr\left(D=1 \mid Z=z'\right)=0.$$

- "Identification at infinity" plays a crucial role throughout the entire literature on policy evaluation.

Imbens Angrist conditions (1994)

- In general, LATE $\neq E(Y_1 - Y_0) = E(\beta)$.
- Not treatment on the treated: $E(\beta \mid D = 1)$.
- Different instruments define different parameters.
- Having a wealth of different strong instruments does not improve the precision of the estimate of any particular parameter (Heckman and Robb, 1986).
- When there are more than two distinct values of Z, Imbens and Angrist use Yitzhaki (1989) weights.

Imbens Angrist conditions (1994)

- Goal of our work: unify literature with a common set of underlying parameters interpretable across studies.
- To understand how to connect the results of various disparate IV estimands within a unified framework.

IV in choice models

$$D = \mathbf{1}\left[D^* > 0\right] \tag{9}$$

$\mathbf{1}[\cdot]$ is an indicator ($\mathbf{1}[A] = 1$ if A true; 0 otherwise).

$$D^* = \mu_D(Z) - V \tag{10}$$

Example: $\mu_D(Z) = \gamma Z$

$$D^* = \gamma Z - V$$

Examples

$$(V \perp\!\!\!\perp Z) \mid X.$$

The propensity score:

$$P(z) = \Pr(D = 1 \mid Z = z) = \Pr(\gamma z > V) = F_V(\gamma z)$$

F_V is the distribution of V.

Examples

Generalized Roy model

$$D = \mathbf{1}[Y_1 - Y_0 - C > 0]$$

$$\begin{aligned} \text{Costs } C &= \mu_C(W) + U_C \\ Z &= (X, W) \\ \mu_D(Z) &= \mu_1(X) - \mu_0(X) - \mu_C(W) \\ V &= -(U_1 - U_0 - U_C). \end{aligned}$$

Heterogeneous response model

In a general model with heterogenous responses, specification of $P(Z)$ and its relationship with the instrument play a crucial role.

$$
\begin{aligned}
\operatorname{Cov}(Z, \eta D) &= E\left(\left(Z-\bar{Z}\right)\eta D\right) \\
&= E\left(\left(Z-\bar{Z}\right)\eta \mid D=1\right)\Pr\left(D=1\right) \\
&= E\Big(\left(Z-\bar{Z}\right)\eta \mid \underbrace{\gamma Z > V}_{\substack{F_V(\gamma Z) > F_V(V) \\ P(Z) > U_D}}\Big)\underbrace{\Pr\left(\gamma Z > V\right)}_{P(Z)}.
\end{aligned}
$$

- Probability of selection enters the covariance even though we use only one component of Z as an instrument.

- Selection models control for this dependence induced by choice.

Selection models

Assume

$$(U_1, U_0, V) \perp\!\!\!\perp Z \tag{11}$$

$$[\text{Alternatively } (\varepsilon, \eta, V) \perp\!\!\!\perp Z].$$

$$\eta = (U_1 - U_0),\ \varepsilon = U_0 \tag{12}$$

$$\begin{aligned} E\left(Y \mid D = 0, Z = z\right) &= E(Y_0 \mid D = 0, Z = z) \\ &= \alpha + E(U_0 \mid \gamma z < V) \end{aligned}$$

$$E(Y \mid D = 0, Z = z) = \alpha + \underbrace{K_0(P(z))}_{\text{control function}}$$

Selection models

$$
\begin{aligned}
E\left(Y \mid D=1, Z=z\right) &= E\left(Y_1 \mid D=1, Z=z\right) \\
&= \alpha + \bar{\beta} + E\left(U_1 \mid \gamma z > V\right) \\
&= \alpha + \bar{\beta} + \underbrace{K_1(P(z))}_{\text{control function}}
\end{aligned}
$$

- $K_0(P(z))$ and $K_1(P(z))$ are control functions in the sense of Heckman and Robb (1985, 1986).
- $P(z)$ is an essential ingredient.
- Matching: $K_1\left(P(z)\right) = K_0\left(P(z)\right)$.

- In a model where β is variable and not independent of V, misspecification of Z affects the interpretation of what IV estimates analogous to its role in selection models.
- Misspecification of Z affects both approaches to identification.
- This is a new phenomenon in models with heterogenous β.

Model for outcomes

$$\begin{aligned} Y_1 &= \mu_1(X, U_1) \\ Y_0 &= \mu_0(X, U_0). \end{aligned} \tag{13}$$

- X are observed and (U_1, U_0) are unobserved by the analyst.
- The X may be dependent on U_0 and U_1.
- Generalize choice model (9) and (10) for D^*, a latent utility.

Model for outcomes

$$D^* = \mu_D(Z) - V \text{ and } D = \mathbf{1}(D^* \geq 0) \tag{14}$$

$\mu_D(Z) - V$ can be interpreted as a net utility for a person with characteristics (Z, V).

- $\beta = Y_1 - Y_0 = \mu_1(X, U_1) - \mu_0(X, U_0)$ (Treatment Effect)

Model for outcomes

- A special case that links our analysis to standard models in econometrics:
- $Y_1 = X\beta_1 + U_1$ and
- $Y_0 = X\beta_0 + U_0$; so
- $\beta = X(\beta_1 - \beta_0) + (U_1 - U_0)$.
- In the case of separable outcomes, heterogeneity in β arises because in general $U_1 \neq U_0$ and people differ in their X.
- Heckman-Vytlacil conditions (1999,2001, 2005)

Assumptions

(A-1)

The distribution of $\mu_D(Z)$ conditional on X is nondegenerate (Rank Condition for IV). This says that we can vary Z (excluded from outcome equations) given X. Key property of an instrument.

(A-2)

(U_0, U_1, V) are independent of Z conditional on X (Independence Condition for IV). Z is not affecting potential outcomes or affecting the unobservables affecting choices.

Assumptions

(A-3)

The distribution of V is continuous (not essential).

(A-4)

$E\,|Y_1| < \infty$, *and* $E\,|Y_0| < \infty$ *(Finite Means).*

Assumptions

(A-5)

$1 > \Pr(D = 1 \mid X) > 0$ *(For each X there is a treatment group and a comparison group).*

(A-6)

Let X_0 denote the counterfactual value of X that would have been observed if D is set to 0. X_1 is defined analogously. Thus $X_d = X$, for $d = 0, 1$ (The X_d are invariant to counterfactual manipulations).

- Separability between V and $\mu_D(Z)$ in choice equation is conventional.
- Plays an important role in the properties of instrumental variable estimators in models with essential heterogeneity.
- It implies monotonicity (uniformity) condition (IV-3) from choice equation (14).
- Vytlacil (2002) shows that independence and monotonicity (IV-3) imply the existence of a V and representation (14) given some regularity conditions.

Use probability integral transform to write

$$D = \mathbf{1}\left[F_V\left(\mu_D\left(Z\right)\right) > F_V\left(V\right)\right] = \mathbf{1}\left[P\left(Z\right) > U_D\right] \tag{15}$$

$U_D = F_V\left(V\right)$ and $P\left(Z\right) = F_V\left(\mu_D(Z)\right) = \Pr[D = 1 \mid Z]$

- $P(Z)$ is transformation of mean scale utility in a discrete choice model.

LATE, the marginal treatment effect and instrumental variables

- A basic parameter that can be used to unify the treatment effect literature:

$$\begin{aligned}\Delta^{\mathrm{MTE}}(x, u_D) &= E(Y_1 - Y_0 \mid X = x, U_D = u_D). \\ &= E(\beta \mid X = x, V = v)\end{aligned}$$

- MTE and the local average treatment effect (LATE) parameter are closely related.
- For $(z, z') \in \mathcal{Z}(x) \times \mathcal{Z}(x)$ so that $P(z) > P(z')$, under (IV-3) and independence (A-2), LATE is:

$$\Delta^{\mathrm{LATE}}(z', z) = E(Y_1 - Y_0 \mid D(z) = 1, D(z') = 0) \qquad (16)$$

LATE can be written in a fashion free of any instrument:

$$\begin{aligned} & E\left(Y_1 - Y_0 \mid D(z) = 1, D(z') = 0\right) \\ & \quad = E\left(Y_1 - Y_0 \mid u_D' < U_D < u_D\right) \\ & \quad = \Delta^{\mathrm{LATE}}(u_D', u_D) \end{aligned} \tag{17}$$

$$u_D = \Pr(D\left(z\right) = 1) = \Pr\left(D\left(z\right) = 1 \mid Z = z\right) = \Pr(D\left(z\right) = 1) = P(z)$$
$$u_D' = \Pr\left(D\left(z'\right) = 1 \mid Z = z'\right) = \Pr(D\left(z'\right) = 1) = P(z')$$

The z just help us define evaluation points for the u_D.

- Under (A-1)–(A-5), all standard treatment parameters are weighted averages of MTE with weights that can be estimated.

Table 1A: treatment effects and estimands as weighted averages of the marginal treatment effect

$$\text{ATE}(x) = E\left(Y_1 - Y_0 \mid X = x\right) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, du_D$$

$$\text{TT}(x) = E\left(Y_1 - Y_0 \mid X = x, D = 1\right) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TT}}(x, u_D)\, du_D$$

$$\text{TUT}(x) = E\left(Y_1 - Y_0 \mid X = x, D = 0\right) = \int_0^1 \Delta^{\text{MTE}}\left(x, u_D\right)\, \omega_{\text{TUT}}\left(x, u_D\right)\, du_D$$

Policy Relevant Treatment Effect (x)

$$= E\left(Y_{a'} \mid X = x\right) - E\left(Y_a \mid X = x\right) = \int_0^1 \Delta^{\text{MTE}}\left(x, u_D\right)\, \omega_{\text{PRTE}}\left(x, u_D\right)\, du_D$$

for two policies a and a' that affect the Z but not the X

$$\text{IV}_J(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{IV}}^J(x, u_D)\, du_D, \text{ given instrument } J$$

$$\text{OLS}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{OLS}}(x, u_D)\, du_D$$

Table 1B: weights

$$\omega_{\text{ATE}}(x, u_D) = 1$$

$$\omega_{\text{TT}}(x, u_D) = \left[\int_{u_D}^{1} f(p \mid X = x) dp\right] \frac{1}{E(P \mid X = x)}$$

$$\omega_{\text{TUT}}\left(x, u_D\right) = \left[\int_0^{u_D} f\left(p | X = x\right) dp\right] \frac{1}{E\left(\left(1 - P\right) | X = x\right)}$$

$$\omega_{\text{PRTE}}(x, u_D) = \left[\frac{F_{P_{a'}, X}(u_D) - F_{P_a, X}(u_D)}{\Delta \overline{P}}\right]$$

Table 1B: weights

$$\omega_{\text{IV}}^{J}(x, u_D) = \frac{\int_{u_D}^{1} (J(Z) - E(J(Z) \mid X = x)) \int f_{J,P|X} (j, t \mid X = x) \; dt \, dj}{\text{Cov}(J(Z), D \mid X = x)}$$

$$\omega_{\text{OLS}}(x, u_D) = 1 + \frac{\left\{ \begin{array}{c} E(U_1 \mid X = x, U_D = u_D) \, \omega_1(x, u_D) \\ -E(U_0 \mid X = x, U_D = u_D) \, \omega_0(x, u_D) \end{array} \right\}}{\Delta^{\text{MTE}}(x, u_D)}$$

Table 1B: weights

$$\omega_1(x, u_D) = \left[\int_{u_D}^{1} f(p \mid X = x)\, dp\right] \left[\frac{1}{E(P \mid X = x)}\right]$$

$$\omega_0(x, u_D) = \left[\int_{0}^{u_D} f(p \mid X = x)\, dp\right] \frac{1}{E((1 - P) \mid X = x)}$$

Source: Heckman and Vytlacil (2005)

Relationships Among Parameters Using the Index Structure

- From the definition $D(z) = \mathbf{1}\left(U_D \leq P(z)\right)$,

$$\Delta^{\mathrm{TT}}(x, P(z)) = E(\Delta | X = x, U_D \leq P(z)). \tag{18}$$

- Consider $\Delta^{\mathrm{LATE}}(x, P(z), P(z'))$.

$$\begin{aligned} & E(Y|X = x, P(Z) = P(z)) \\ & \quad = P(z)\Big[E(Y_1|X = x, P(Z) = P(z), D = 1)\Big] \\ & \qquad + (1 - P(z))\Big[E\left(Y_0|X = x, P(Z) = P(z), D = 0\right)\Big] \\ & \quad = \int_0^{P(z)} E(Y_1|X = x, U_D = u_D) du_D + \int_{P(z)}^1 E(Y_0|X = x, U_D = u_D) du_D. \end{aligned}$$

- So that

$$E(Y|X=x,P(Z)=P(z))-E(Y|X=x,P(Z)=P(z'))$$
$$=\int_{P(z')}^{P(z)}E(Y_1|X=x,U_D=u_D)du_D-\int_{P(z')}^{P(z)}E(Y_0|X=x,U_D=u_D)du_D,$$

and thus

$$\Delta^{\text{LATE}}(x,P(z),P(z'))=E(\Delta|X=x,P(z')\leq U_D\leq P(z)).$$

- Notice that this expression could be taken as an alternative definition of LATE.
- Note that in this expression we could replace $P(z)$ and $P(z')$ with u_D and u_D'.
- No instrument needs to be available to define LATE.

- Rewrite these relationships in succinct form:

$$\Delta^{\text{MTE}}(x, u_D) = E(\Delta | X = x, U_D = u_D) \tag{19}$$

$$\Delta^{\text{ATE}}(x) = \int_0^1 E(\Delta | X = x, U_D = u_D) du_D$$

$$P(z)[\Delta^{\text{TT}}(x, P(z))] = \int_0^{P(z)} E(\Delta | X = x, U_D = u_D) du_D$$

$$(P(z) - P(z'))[\Delta^{\text{LATE}}(x, P(z), P(z'))] = \int_{P(z')}^{P(z)} E(\Delta | X = x, U_D = u_D) du_D$$

- Everywhere in these expressions can replace $P(z)$ with u_D and $P(z')$ with u'_D.
- Each parameter is an average value of MTE, $E(\Delta \mid X = x, U_D = u_D)$, but for values of U_D lying in different intervals and with different weighting functions.
- MTE defines the treatment effect more finely than do LATE, ATE, or TT.
- The relationship between MTE and LATE or TT conditional on $P(z)$ is analogous to the relationship between a probability density function and a cumulative distribution function.

- The probability density function and the cumulative distribution function represent the same information, but for some purposes the density function is more easily interpreted.
- Likewise, knowledge of TT for all $P(z)$ evaluation points is equivalent to knowledge of the MTE for all u evaluation points, so it is not the case that knowledge of one provides more information than knowledge of the other.
- However, in many choice-theoretic contexts it is often easier to interpret MTE than the TT or LATE parameters.
- It has the interpretation as a measure of willingness to pay on the part of people on a specified margin of participation in the program.

- $\Delta^{\text{MTE}}(x, u_D)$ is the average effect for people who are just indifferent between participation in the program $(D = 1)$ or not $(D = 0)$ if the instrument is externally set so that $P(Z) = u_D$.
- For values of u_D close to zero, $\Delta^{\text{MTE}}(x, u_D)$ is the average effect for individuals with unobservable characteristics that make them the most inclined to participate in the program $(D = 1)$, and for values of u_D close to one it is the average treatment effect for individuals with unobserved (by the econometrician) characteristics that make them the least inclined to participate.

- ATE integrates $\Delta^{\text{MTE}}(x, u_D)$ over the entire support of U_D (from $u_D = 0$ to $u_D = 1$).
- It is the average effect for an individual chosen at random from the entire population.

- $\Delta^{\text{TT}}(x, P(z))$ is the average treatment effect for persons who chose to participate at the given value of $P(Z) = P(z)$; it integrates $\Delta^{\text{MTE}}(x, u_D)$ up to $u_D = P(z)$.
- As a result, it is primarily determined by the MTE parameter for individuals whose unobserved characteristics make them the most inclined to participate in the program.
- LATE is the average treatment effect for someone who would not participate if $P(Z) \leq P(z')$ and would participate if $P(Z) \geq P(z)$.
- The parameter $\Delta^{\text{LATE}}(x, P(z), P(z'))$ integrates $\Delta^{\text{MTE}}(x, u_D)$ from $u_D = P(z')$ to $u_D = P(z)$.

- Using the third expression in equation (19) to substitute into equation (18), we obtain an alternative expression for the TT parameter as a weighted average of MTE parameters:

$$\Delta^{\mathrm{TT}}(x) = \int_0^1 \frac{1}{p}\left[\int_0^p E(\Delta|X = x, U_D = u_D)du_D\right] dF_{P(Z)|X,D}(p|x, D = 1).$$

- Using Bayes' rule, it follows that

$$dF_{P(Z)|X,D}(p|x, 1) = \frac{\Pr(D = 1|X = x, P(Z) = p)}{\Pr(D = 1|X = x)} dF_{P(Z)|X}(p|x).$$

- Since $\Pr(D = 1|X = x, P(Z) = p) = p$, it follows that

$$\Delta^{\mathrm{TT}}(x) \tag{20}$$
$$= \frac{1}{\Pr(D = 1|X = x)} \int_0^1 \left(\int_0^p E(\Delta|X = x, U_D = u_D) du_D \right) dF_{P(Z)|X}(p|x).$$

- Note further that since $\Pr(D = 1|X = x) = E(P(Z)|X = x) = \int_0^1 (1 - F_{P(Z)|X}(t|x))dt$, we can reinterpret (20) as a weighted average of local IV parameters where the weighting is similar to that obtained from a length-biased, size-biased, or *P*-biased sample.

$$
\begin{aligned}
&\Delta^{\mathrm{TT}}(x) \\
&= \frac{1}{\Pr(D=1|X=x)} \\
&\quad \cdot \int_0^1 \left(\int_0^1 \mathbf{1}(u_D \le p) E(\Delta|X=x, U_D=u_D) du_D \right) dF_{P(Z)|X}(p|x) \\
&= \frac{1}{\int (1 - F_{P(Z)|X}(t|x)) dt} \\
&\quad \int_0^1 \left(\int_0^1 E(\Delta|X=x, U_D=u_D) \mathbf{1}(u_D \le p) dF_{P(Z)|X}(p|x) \right) du_D \\
&= \int_0^1 E(\Delta|X=x, U_D=u_D) \left(\frac{1 - F_{P(Z)|X}(u_D|x)}{\int (1 - F_{P(Z)|X}(t|x)) dt} \right) du_D \\
&= \int_0^1 E(\Delta|X=x, U_D=u_D) g_x(u_D) du_D
\end{aligned}
$$

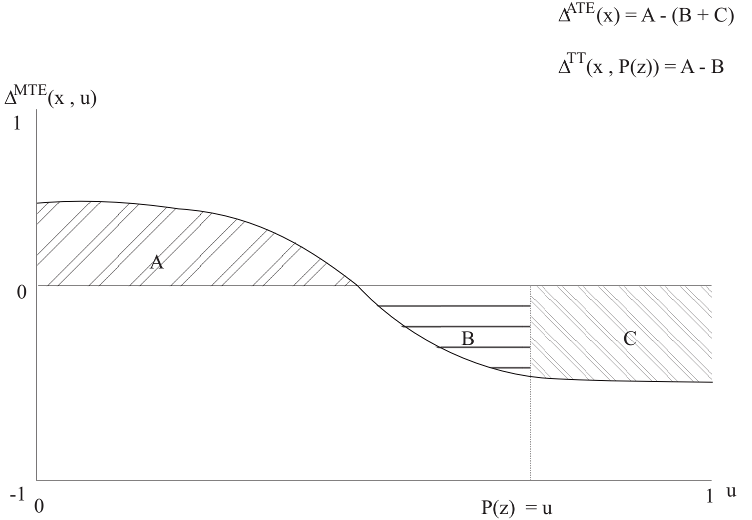
where $g_x(u_D) = \frac{1 - F_{P(Z)|X}(u_D|x)}{\int (1 - F_{P(Z)|X}(t|x)) dt}$.

- Thus $g_x(u_D)$ is a *weighted distribution* (Rao, 1985).
- Since $g_x(u_D)$ is a nonincreasing function of u_D, we have that drawings from $g_x(u_D)$ oversample persons with low values of U_D, i.e., values of unobserved characteristics that make them the most likely to participate in the program no matter what their value of $P(Z)$.
- Since
$$\Delta^{\text{MTE}}(x, u_D) = E(\Delta | X = x, U_D = u_D)$$
it follows that
$$\Delta^{\text{TT}}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D) g_x(u_D) du_D.$$

- The TT parameter is thus a weighted version of MTE, where $\Delta^{\text{MTE}}(x, u_D)$ is given the largest weight for low u values and is given zero weight for $u_D \geq p_x^{max}$, where p_x^{max} is the maximum value in the support of $P(Z)$ conditional on $X = x$.

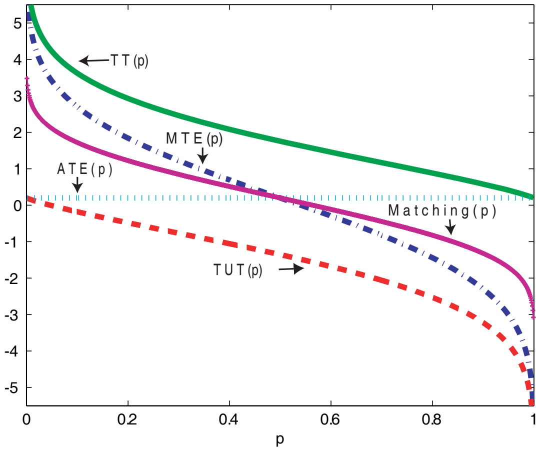
- Figure A-1 graphs the relationship between $\Delta^{\text{MTE}}(u_D)$, Δ^{ATE} and $\Delta^{\text{TT}}(P(z))$, assuming that the gains are the greatest for those with the lowest U_D values and that the gains decline as U_D increases.
- The curve is the MTE parameter as a function of u_D, and is drawn for the special case where the outcome variable is binary so that MTE parameter is bounded between -1 and 1.
- The ATE parameter averages $\Delta^{\text{MTE}}(u_D)$ over the full unit interval (i.e. is the area under A minus the area under B and C in the figure).

Figure A-1. MTE Integrates to ATE and TT Under Full Support (for dichotomous outcome)



Source: Heckman and Vytlacil (2000).

Figure 9: treatment parameters and OLS matching as a function of $P(Z) = p$



- $\Delta^{\text{TT}}(P(z))$ averages $\Delta^{\text{MTE}}(u_D)$ up to the point $P(z)$ (is the area under A minus the area under B in the figure).
- Because $\Delta^{\text{MTE}}(u_D)$ is assumed to be declining in u, the TT parameter for any given $P(z)$ evaluation point is larger then the ATE parameter.

- Equation (19) relates each of the other parameters to the MTE parameter.
- One can also relate each of the other parameters to the LATE parameter.
- This relationship turns out to be useful later on in this chapter when we encounter conditions where LATE can be identified but MTE cannot.
- MTE is the limit form of LATE:

$$\Delta^{\mathrm{MTE}}(x,p)=\lim_{p'\to p}\Delta^{\mathrm{LATE}}(x,p,p').$$

- Direct relationships between LATE and the other parameters are easily derived.
- The relationship between LATE and ATE is immediate:

$$\Delta^{\text{ATE}}(x) = \Delta^{\text{LATE}}(x, 0, 1).$$

- Using Bayes' rule, the relationship between LATE and TT is

$$\Delta^{\text{TT}}(x) = \int_0^1 \Delta^{\text{LATE}}(x, 0, p) \frac{p}{\Pr(D = 1 | X = x)} dF_{P(Z)|X}(p|x). \tag{21}$$

Derivation of PRTE and Implications of Noninvariance for PRTE

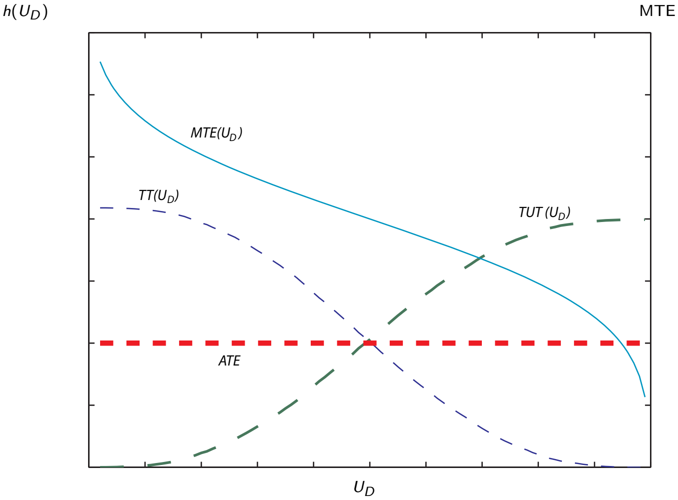
$$
\begin{aligned}
E\left(Y_p \mid X\right) &= \int_0^1 E(Y_p|X, P_p(Z_p) = t)\, dF_{P_p|X}(t) \\
&= \int_0^1 \left[\int_0^1 [\mathbf{1}_{[0,t]}(u_D)E(Y_{1,p} \mid X, U_D = u_D) \right. \\
&\qquad \left. + \mathbf{1}_{(t,1]}(u_D)E(Y_{0,p} \mid X, U_D = u_D)]\, du\right]\, dF_{P_p|X}(t) \\
&= \int_0^1 \left[\int_0^1 [\mathbf{1}_{[u_D,1]}(t)E(Y_{1,p} \mid X, U_D = u_D) \right. \\
&\qquad \left. + \mathbf{1}_{(0,u_D]}(t)E(Y_{0,p} \mid X, U_D = u_D)]\, dF_{P_p|X}(t)\right]\, du_D \\
&= \int_0^1 \left[(1 - F_{P_p|X}(u_D))E(Y_{1,p} \mid X, U_D = u_D) \right. \\
&\qquad \left. + F_{P_p|X}(u_D)E(Y_{0,p} \mid X, U_D = u_D)\right]\, du_D .
\end{aligned}
$$

- This derivation involves changing the order of integration.
- Note that from (A-4),

$$E\left|\mathbf{1}_{[0,t]}(u_D)E(Y_{1,p} \mid X, U_D = u_D) + \mathbf{1}_{(t,1]}(u_D)E(Y_{0,p} \mid X, U_D = u_D)\right| \\ \leq E(|Y_1| + |Y_0|) < \infty,$$

so the change in the order of integration is valid by Fubini's theorem.

Figure 2: weights for the marginal treatment effect for different parameters



- $E(\beta \mid U_D = u_D)$ does not vary with u_D.
- "Standard case."
- ATE = TT = LATE = policy counterfactuals = plim IV.

When will $E(\beta \mid U_D = u_D)$ not vary with u_D?

1. If $U_1 = U_0 \Rightarrow \beta$ a Constant.
2. More Generally, if $U_1 - U_0$ is mean independent of U_D, so treatment effect heterogeneity is allowed but individuals do not act upon their own idiosyncratic effect.

Consider standard analysis.

$$\ln Y = \alpha + (\bar{\beta} + U_1 - U_0)D + U_0$$

plim of OLS:

$$\begin{aligned}
& E(\ln Y \mid D=1) - E(\ln Y \mid D=0) \\
&= \bar{\beta} + E(U_1 - U_0 \mid D=1) + \left\{ \begin{array}{l} E(U_0 \mid D=1) \\ \quad -E(U_0 \mid D=0) \end{array} \right\} \\
&= \underbrace{\text{ATE} + \text{Sorting Gain}} + \text{Ability Bias} \\
&= \text{TT} + \text{Ability Bias}
\end{aligned}$$

- If ATE is a parameter of interest, OLS suffers from both sorting bias and ability bias.
- If TT is parameter of interest, OLS suffers from ability bias.
- Using IV removes ability bias, but changes the parameter being estimated (neither ATE nor TT in general).
- Different IV Weight MTE differently.
- We derive IV weights below.

- $\therefore$ IV Instrument Dependent (which Z used and which values of Z used).
- Hence studies using different Z are not comparable.
- How to make studies comparable?
- We can test to see if these complications are required in any particular empirical analysis.

Testing for essential heterogeneity

$$
\begin{aligned}
E\left(Y \mid Z=z\right) &= E(Y \mid P(Z)=p) \text{ (index sufficiency)} \\
&= E\left(DY_1+(1-D)Y_0 \mid P\left(Z\right)=p\right) \\
&= E(Y_0)+E\left(D\left(Y_1-Y_0\right)|P\left(Z\right)=p\right) \\
&= E(Y_0)+\left[\begin{array}{r} E(Y_1-Y_0|D=1,P\left(Z\right)=p) \\ \cdot \Pr\left(D=1 \mid Z=z\right) \end{array}\right] \\
&= E(Y_0)+\int_0^p E(Y_1-Y_0|U_D=u_D)\,du_D.
\end{aligned}
$$

Testing for essential heterogeneity

As a consequence, we get LIV (Local Instrumental Variables), which identifies MTE

$$\underbrace{\frac{\partial}{\partial P(z)} E\left(Y \mid Z = z\right)\bigg|_{P(Z)=u_D}}_{LIV} = \underbrace{E(Y_1 - Y_0 | U_D = u_D)}_{MTE}. \tag{22}$$

- When $\beta \perp\!\!\!\perp D$, Y is linear in $P\left(Z\right)$:

  $$E\left(Y \mid Z\right) = a + bP\left(Z\right) \tag{23}$$

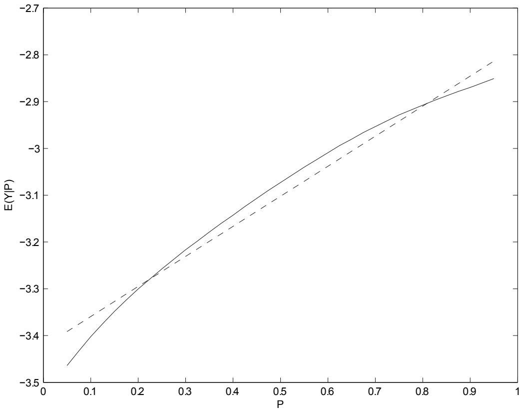
  where $b = \Delta^{\text{MTE}} = \Delta^{\text{ATE}} = \Delta^{\text{TT}}$.

- These results are valid whether or not Y_1 and Y_0 are separable in U_1 and U_0.

- Therefore we can identify the treatment parameters using estimated weights and estimated MTE.

Example: college attendance on wages for high school graduates

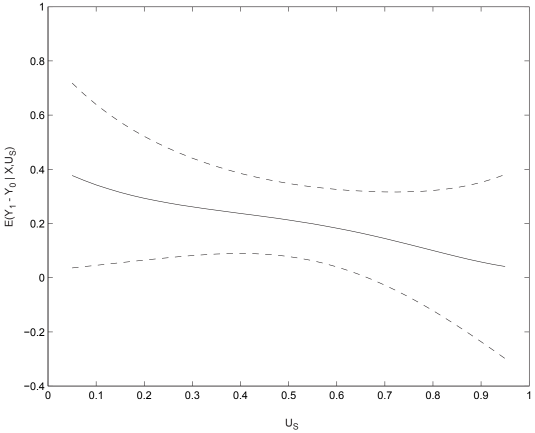
$E(Y \mid X, P)$ as a function of P for average X



Source: Carneiro, Heckman and Vytlacil (2006)

Example: college attendance on wages for high school graduates

$E(Y_1 - Y_0 \mid X, U_S)$ estimated using locally quadratic regression (averaged over X)



Source: Carneiro, Heckman and Vytlacil (2006)

Understanding what linear IV estimates

- Consider $J(Z)$ as an instrument, a scalar function of Z.

$$\Delta_J^{\mathrm{IV}} = \frac{\mathrm{Cov}(Y, J(Z))}{\mathrm{Cov}(D, J(Z))}$$

- Express it as a weighted average of MTE.
- Z can be a vector of instruments.

Digression: Yitzhaki's theorem and extensions

Theorem

Assume (Y, X) *i.i.d.* $\quad E(|Y|) < \infty \quad E(|X|) < \infty$

$$\mu_Y = E(Y) \qquad \mu_X = E(X)$$

$E(Y \mid X) = g(X)$
Assume $g'(X)$ *exists and* $E\left(|g'(X)|\right) < \infty$.

Yitzhaki's theorem

Theorem (cont.)

Then,

$$\frac{\mathrm{Cov}(Y, X)}{Var(X)} = \int_{-\infty}^{\infty} g'(t)\, \omega(t)\, dt,$$

where

$$\begin{aligned} \omega(t) &= \frac{1}{Var(X)} \int_{t}^{\infty} (x - \mu_X)\, f_X(x)\, dx \\ &= \frac{1}{Var(X)} E\left(X - \mu_X \mid X > t\right) \Pr\left(X > t\right). \end{aligned}$$

$$Y = \pi X + \eta,$$

$$\pi = \frac{\mathrm{Cov}(Y, X)}{Var(X)}.$$

Proof of Yitzhaki's theorem

Proof.

$$\begin{aligned} \mathrm{Cov}(Y, X) &= \mathrm{Cov}\left(E(Y \mid X), X\right) = \mathrm{Cov}\left(g(X), X\right) \\ &= \int_{-\infty}^{\infty} g(t)(t - \mu_X)\, f_X(t)\, dt \end{aligned}$$

where t is an argument of integration.

Proof of Yitzhaki's theorem

cont.

Integration by parts:

$$\begin{aligned}
\mathrm{Cov}(Y,X) &= g(t)\int_{-\infty}^{t}(x-\mu_X)\,f_X(x)\,dx\Big|_{-\infty}^{\infty} \\
&\qquad - \int_{-\infty}^{\infty} g'(t)\int_{-\infty}^{t}(x-\mu_X)\,f_X(x)\,dx\,dt \\
&= \int_{-\infty}^{\infty} g'(t)\int_{t}^{\infty}(x-\mu_X)\,f_X(x)\,dx\,dt,
\end{aligned}$$

since $E\left(X-\mu_X\right)=0$.

Proof of Yitzhaki's theorem

cont.

Therefore,

$$\mathrm{Cov}(Y, X) = \int_{-\infty}^{\infty} g'(t)\, E\left(X - \mu_X \mid X > t\right) \Pr\left(X > t\right)\, dt.$$

$\therefore$ Result follows with

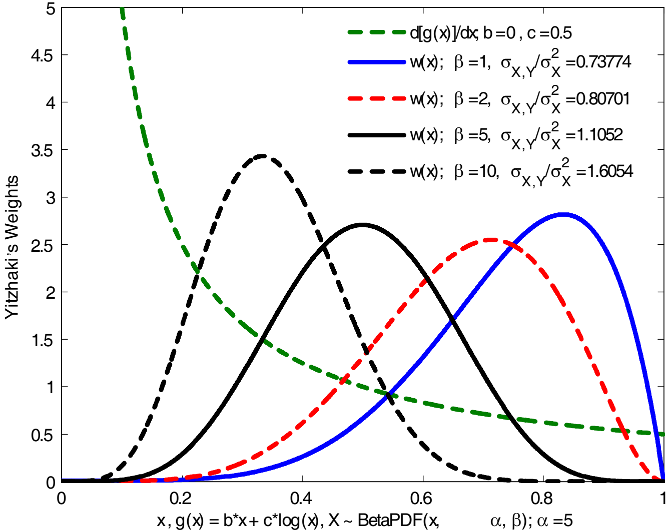
$$\omega(t) = \frac{1}{Var(X)} E\left(X - \mu_X \mid X > t\right) \Pr\left(X > t\right)$$

$\square$

- Weights positive.
- Integrate to one (use integration by parts formula).
- $= 0$ when $t \to \infty$ and $t \to -\infty$.
- Weight reaches its peak at $t = \mu_X$, if f_X has density at $x = \mu_X$:

$$\begin{aligned} \frac{d}{dt} \int_t^{\infty} (x - \mu_X)\, f_X(x)\, dx\, dt &= -(t - \mu_X) f_X(t) \\ &= 0 \quad \text{at } t = \mu_X. \end{aligned}$$

Yitzhaki's weights for $X \sim \text{BetaPDF}(x, \alpha, \beta)$



Yitzhaki's weights for $X \sim \text{BetaPDF}(x, \alpha, \beta)$

$$E\left(Y|X=x\right) = g(x) \Rightarrow \frac{Cov(X,Y)}{Var(X)} = \int_{-\infty}^{\infty} g'(t)w(t)dx$$

$$w(t) = \frac{1}{Var(X)} E\left(X|X>t\right) \cdot \Pr\left(X>t\right)$$

$$\mathbf{X} \sim BetaPDF(x,\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)};\ \alpha = 5;$$

$$\mathbf{g(x) = 0.5 \cdot x + 0.5} \cdot \log(\mathbf{X})$$

- Can apply Yitzhaki's analysis to the treatment effect model

$$Y = \alpha + \beta D + \varepsilon$$

- $P(Z)$, the propensity score is the instrument:

$$E\left(Y \mid Z = z\right) = E\left(Y \mid P(Z) = p\right)$$

$$
\begin{aligned}
E\left(Y \mid P(Z)=p\right) &= \alpha + E\left(\beta D \mid P(Z)=p\right) \\
&= \alpha + E\left(\beta \mid D=1, P(Z)=p\right) p \\
&= \alpha + E\left(\beta \mid P(Z)>U_D, P(Z)=p\right) p \\
&= \alpha + E\left(\beta \mid p>U_D\right) p \\
&= \alpha + \underbrace{\int \beta \int_0^p f\left(\beta, u_D\right)\, du_D}_{g(p)}
\end{aligned}
$$

- Derivative with respect to p is MTE.
- $g'(p) =$MTE and weights as before.

- Under uniformity,

$$\begin{aligned} \frac{\partial E\left(Y \mid P\left(Z\right)=p\right)}{\partial p} &= E\left(Y_1 - Y_0 \mid U_D = u_D\right) \\ &= \Delta^{MTE}\left(u_D\right). \end{aligned}$$

- More generally, it is LIV $= \frac{\partial E(Y|P(Z)=p)}{\partial p}$.
- Yitzhaki's result does not rely on uniformity; true of any regression of Y on P.
- Estimates a weighted net effect.
- The expression can be generalized.
- It produces Heckman-Vytlacil weights.

The Heckman-Vytlacil weight as a Yitzhaki weight

Proof.

$$\begin{aligned}\mathrm{Cov}\left(J\left(Z\right),Y\right) &= E\left(Y\cdot\widetilde{J}\right)=E\left(E\left(Y\mid Z\right)\cdot\widetilde{J}(Z)\right)\\ &= E\left(E\left(Y\mid P\left(Z\right)\right)\cdot\widetilde{J}(Z)\right)\\ &= E\left(g\left(P\left(Z\right)\right)\cdot\widetilde{J}(Z)\right).\end{aligned}$$

$$\widetilde{J}=J\left(Z\right)-E\left(J\left(Z\right)\mid P\left(Z\right)\geq u_D\right),$$
$$E\left(Y\mid P\left(Z\right)\right)=g\left(P\left(Z\right)\right).$$

The Heckman-Vytlacil weight as a Yitzhaki weight

cont.

$$\begin{aligned}
\operatorname{Cov}(J(Z), Y) &= \int_0^1 \int_{\underline{J}}^{\overline{J}} g(u_D) \tilde{j} f_{P,J}(u_D, j) \, \mathrm{d}j \mathrm{d}u_D \\
&= \int_0^1 g(u_D) \int_{\underline{J}}^{\overline{J}} \tilde{j} f_{P,J}(u_D, j) \, \mathrm{d}j \mathrm{d}u_D.
\end{aligned}$$

The Heckman-Vytlacil weight as a Yitzhaki weight

cont.

Use integration by parts:

$$
\begin{aligned}
&\mathrm{Cov}\left(J\left(Z\right), Y\right) \\
&= g\left(u_D\right) \int_0^{u_D} \int_{\underline{J}}^{\overline{J}} \widetilde{j} f_{P,J}\left(p, j\right) \, \mathrm{d}j \mathrm{d}p \bigg|_0^1 \\
&\qquad - \int_0^1 g'\left(u_D\right) \int_0^{u_D} \int_{\underline{J}}^{\overline{J}} \widetilde{j} f_{P,J}\left(p, j\right) \, \mathrm{d}j \mathrm{d}p \mathrm{d}u_D \\
&= \int_0^1 g'\left(u_D\right) \int_{u_D}^1 \int_{\underline{J}}^{\overline{J}} \widetilde{j} f_{P,J}\left(p, j\right) \, \mathrm{d}j \mathrm{d}p \mathrm{d}u_D \\
&= \int_0^1 g'\left(u_D\right) E\left(\widetilde{J}(Z) \mid P\left(Z\right) \geq u_D\right) \Pr\left(P\left(Z\right) \geq u_D\right) \mathrm{d}u_D.
\end{aligned}
$$

The Heckman-Vytlacil weight as a Yitzhaki weight

cont.

Thus:

$$g'(u_D) = \left. \frac{\partial E(Y \mid P(Z) = p)}{\partial P(Z)} \right|_{p=u_D} = \Delta^{\text{MTE}}(u_D).$$

□

- Under our assumptions the Yitzhaki weights and ours are equivalent.
- $$\begin{aligned}&\operatorname{Cov}\left(J\left(Z\right), Y\right) \qquad (24)\\&= \int_0^1 \Delta^{\text{MTE}}(u_D) E(J(Z) - E(J(Z)) \mid P(Z) \geq u_D) \Pr(P(Z) \geq u_D) du_D.\end{aligned}$$
- Using (24),

$$\begin{aligned}\operatorname{Cov}\left(J\left(Z\right), Y\right) &= E\left(Y \cdot \tilde{J}\right) = E\left(E\left(Y \mid Z\right) \cdot \tilde{J}(Z)\right)\\&= E\left(E\left(Y \mid P\left(Z\right)\right) \cdot \tilde{J}(Z)\right)\\&= E\left(g\left(P\left(Z\right)\right) \cdot \tilde{J}(Z)\right).\end{aligned}$$

- The third equality follows from index sufficiency and $\tilde{J} = J(Z) - E(J(Z) \mid P(Z) \geq u_D)$, where $E(Y \mid P(Z)) = g(P(Z))$.

- Writing out the expectation and assuming that $J(Z)$ and $P(Z)$ are continuous random variables with joint density $f_{P,J}$ and that $J(Z)$ has support $\left[\underline{J}, \overline{J}\right]$,

$$
\begin{aligned}
\operatorname{Cov}(J(Z), Y) &= \int_0^1 \int_{\underline{J}}^{\overline{J}} g(u_D)\, \tilde{j} f_{P,J}(u_D, j)\, \mathrm{d}j \mathrm{d}u_D \\
&= \int_0^1 g(u_D) \int_{\underline{J}}^{\overline{J}} \tilde{j} f_{P,J}(u_D, j)\, \mathrm{d}j \mathrm{d}u_D .
\end{aligned}
$$

- Using an integration by parts argument as in Yitzhaki (1989) and as summarized in Heckman, Urzua, Vytlacil (2006), we obtain

$$
\begin{aligned}
&\operatorname{Cov}\left(J\left(Z\right), Y\right) \\
&\quad = g\left(u_D\right) \int_0^{u_D} \int_{\underline{J}}^{\overline{J}} \tilde{j} f_{P,J}\left(p, j\right) \mathrm{d}j \mathrm{d}p \bigg|_0^1 \\
&\qquad\qquad - \int_0^1 g'\left(u_D\right) \int_0^{u_D} \int_{\underline{J}}^{\overline{J}} \tilde{j} f_{P,J}\left(p, j\right) \; \mathrm{d}j \mathrm{d}p \mathrm{d}u_D \\
&\quad = \int_0^1 g'\left(u_D\right) \int_{u_D}^1 \int_{\underline{J}}^{\overline{J}} \tilde{j} f_{P,J}\left(p, j\right) \mathrm{d}j \mathrm{d}p \mathrm{d}u_D \\
&\quad = \int_0^1 g'\left(u_D\right) E\left(\tilde{J}\left(Z\right) \mid P\left(Z\right) \geq u_D\right) \Pr\left(P\left(Z\right) \geq u_D\right) \mathrm{d}u_D,
\end{aligned}
$$

which is then exactly the expression given in (24), where

$$
g'\left(u_D\right) = \frac{\partial E\left(Y \mid P\left(Z\right) = p\right)}{\partial P\left(Z\right)} \bigg|_{p=u_D} = \Delta^{\mathrm{MTE}}\left(u_D\right).
$$

Under (A-1)–(A-5) and separable choice model

$$\Delta_J^{IV} = \int_0^1 \Delta^{MTE}(u_D)\, \omega_{IV}^J(u_D)\, du_D \tag{25}$$

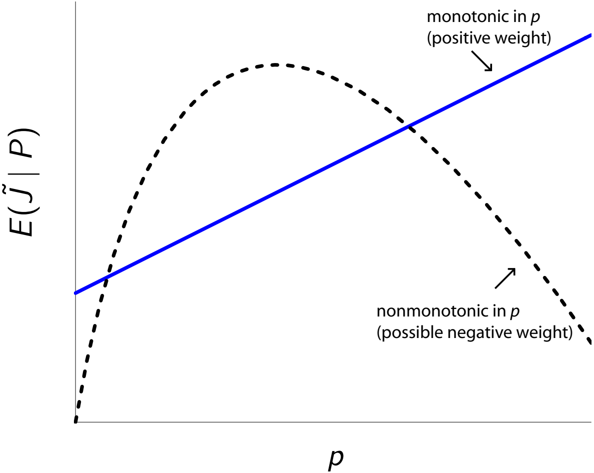
$$\omega_{IV}^J(u_D) = \frac{E\left(J(Z) - \bar{J}(Z) \mid P(Z) > u_D\right) \Pr\left(P(Z) > u_D\right)}{\operatorname{Cov}\left(J(Z), D\right)}. \tag{26}$$

$J(Z)$ and $P(Z)$ do not have to be continuous random variables.

Functional forms of $P(Z)$ and $J(Z)$ are general.

- Dependence between $J(Z)$ and $P(Z)$ gives shape and sign to the weights.
- If $J(Z) = P(Z)$, then weights obviously non-negative.
- If $E(J(Z) - \bar{J}(Z) \mid P(Z) \geq u_D)$ not monotonic in u_D, weights can be negative.

$$\tilde{J} = J - E(J)$$



Therefore, with positive (or negative) regression, can get negative IV weight.

When $J(Z) = P(Z)$, weight (26) follows from Yitzhaki (1989).

- He considers a regression function $E(Y \mid P(Z) = p)$.
- Linear regression of Y on P identifies

$$\beta_{Y,P} = \int_0^1 \left[\frac{\partial E(Y \mid P(Z) = p)}{\partial p} \right] \omega(p) \, dp,$$

$$\omega(p) = \frac{\int_p^1 (t - E(P)) \, dF_P(t)}{Var(P)}.$$

- This is the weight (26) when P is the instrument.
- This expression **does not** require uniformity or monotonicity for the model; consistent with 2-way flows.

Understanding the structure of the IV weights

Recapitulate:

$$\Delta_{\mathrm{IV}}^{J} = \int \Delta^{\mathrm{MTE}}(u_D)\, \omega_{\mathrm{IV}}^{J}(u_D)\, du_D$$

$$\omega_{\mathrm{IV}}^{J}(u_D) = \frac{\int (j - E(J(Z))) \int_{u_D}^{1} f_{J,P}(j,t)\ dt\, dj}{Cov(J(Z), D)} \tag{27}$$

- The weights are always positive if $J(Z)$ is monotonic in the scalar Z.
- In this case $J(Z)$ and $P(Z)$ have the same distribution and $f_{J,P}(j,t)$ collapses to a single distribution.

- The possibility of negative weights arises when $J(Z)$ is not a monotonic function of $P(Z)$.
- It can also arise when there are two or more instruments, and the analyst computes estimates with only one instrument or a combination of the Z instruments that is not a monotonic fuction of $P(Z)$ so that $J(Z)$ and $P(Z)$ are not perfectly dependent.

- The weights can be constructed from data on (J, P, D).
- Data on $(J(Z), P(Z))$ pairs and $(J(Z), D)$ pairs (for each X value) are all that is required.

Discrete instruments $J(Z)$

Discrete Case

- Support of the distribution of $P(Z)$ contains a finite number of values $p_1 < p_2 < \cdots < p_K$.
- Support of the instrument $J(Z)$ is also discrete, taking I distinct values.
- $E(J(Z)|P(Z) \geq u_D)$ is constant in u_D for u_D within any $(p_\ell, p_{\ell+1})$ interval, and $\Pr(P(Z) \geq u_D)$ is constant in u_D for u_D within any $(p_\ell, p_{\ell+1})$ interval.
- Let λ_ℓ denote the weight on the LATE for the interval $(p_\ell, p_{\ell+1})$.

Discrete instruments $J(Z)$

- Under monotonicity, or uniformity

$$\begin{aligned} \Delta_J^{\mathrm{IV}} &= \int E(Y_1 - Y_0 | U_D = u_D)\omega_{\mathrm{IV}}^J(u_D)\, du_D \qquad (28) \\ &= \sum_{\ell=1}^{K-1} \lambda_\ell \int_{p_\ell}^{p_{\ell+1}} E(Y_1 - Y_0 | U_D = u_D) \frac{1}{(p_{\ell+1} - p_\ell)} du_D \\ &= \sum_{\ell=1}^{K-1} \Delta^{\mathrm{LATE}}(p_\ell, p_{\ell+1})\lambda_\ell. \end{aligned}$$

Discrete instruments $J(Z)$

Let j_i be the i^{th} smallest value of the support of $J(Z)$.

$$\lambda_\ell = \frac{\sum_{i=1}^{I} \left(j_i - E(J(Z))\right) \sum_{t>\ell}^{K} \left(f\left(j_i, p_t\right)\right)}{Cov\left(J\left(Z\right), D\right)} \left(p_{\ell+1} - p_\ell\right) \tag{29}$$

Discrete instruments $J(Z)$

- In general, this formula is true, under index sufficiency even if monotonicity is violated.
- It's certainly true under (A-1)–(A-5).
- True where $\Delta^{LATE}(p_\ell, p_{\ell+1})$ is replaced by the Wald estimator, based on $P(z_\ell)$, $\ell = 1, \dots, L$, instruments.
- Observe, LATE here defined in terms of $P(Z)$, the "natural" instrument.

Discrete instruments $J(Z)$

- Generalizes the expression presented by Imbens and Angrist (1994) and Yitzhaki (1989, 1996)
- Their analysis of the case of vector Z only considers the case where $J(Z)$ and $P(Z)$ are perfectly dependent because $J(Z)$ is a monotonic function of $P(Z)$.
- More generally, the weights can be positive or *negative* for any ℓ but they must sum to 1 over the ℓ.

The central role of the propensity score

- For the IV weight to be correctly constructed and interpreted, we need to know the correct model for $P(Z)$.
- IV depends on:
  1. the choice of the instrument $J(Z)$,
  2. its dependence with $P(Z)$,
  3. the specification of the propensity score (i.e., what variables go into Z).
- "Structural" LATE or MTE identified by $P(Z)$.
- Can derive all other instrumental variable estimators in terms of weighted averages of MTE or LATE.

Monotonicity, uniformity and conditional instruments

- Monotonicity or uniformity condition (IV-3) rules out general heterogeneous responses to treatment choices in response to changes in Z.
- The recent literature on instrumental variables with heterogeneous responses is asymmetric.
- The uniformity condition can be violated even when all components of γ are of the same sign if Z is a vector and γ is a nondegenerate random variable.

$$D = \mathbf{1}\left[\gamma Z > \gamma\right]$$

- Uniformity is a condition on a vector.
- Changing one coordinate of Z, holding the other coordinates at different values across people, will not necessarily produce uniformity.
- Let $\mu_D(z) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_1 z_2$, where γ_0, γ_1, γ_2 and γ_3 are constants.
- Consider changing z_1 from a common base state while holding z_2 fixed at different values across people.
- If $\gamma_3 < 0$ then $\mu_D(z)$ does not necessarily satisfy the uniformity condition.

- Positive weights and uniformity are distinct issues.
- Under uniformity, and assumptions (A-1)–(A-5), the weights on MTE or LIV for any particular instrument may be positive or negative.

- If we condition on $Z_2 = z_2, \ldots, Z_K = z_K$ using Z_1 as an instrument, then uniformity is satisfied.
- Effectively convert the problem back to that of a scalar instrument where the weights must be positive.
- The concept of conditioning on other instruments to produce positive weights for the selected instrument is a new idea.

Monotonicity and weights

- Monotonicity is a property needed to get treatment effects with just two values of Z, $Z = z_1$ and $Z = z_2$, to guarantee that IV estimates a treatment effect.
- With multiple values of Z we need to weight to produce linear IV.
- If our IV shifts $P(Z)$ in same way for everyone, it shifts D in the same way for everyone,

$$D = \mathbf{1}\left[P(Z) \geq U_D\right].$$

- If $P(Z)$ is instrument, monotonicity is obviously satisfied.
- If $J(Z)$ is an instrument and not a monotonic function of $P(Z)$, may not shift $P(Z)$ in same way for all people.
- We can get two-way flows if, e.g., we use only one Z or else have a random coefficient model,

$$D = \mathbf{1}\left[\gamma Z \geq V\right].$$

- Negative weights are a tip off of two-way flows.

- If we do not want a treatment effect, who cares?
- We do not always want a treatment effect.
- Go back to ask “What economic question am I trying to answer?”

Treatment effects vs. policy effects

- Even if uniformity condition (IV-3) fails, IV may answer relevant policy questions.
- IV or TSLS estimates a weighted average of marginal responses which may be pointwise positive or negative.
- Policies may induce some people to switch into and others to switch out of choices.
- Net effects are sometimes of interest in many policy analyses.

- Thus, subsidized housing in a region supported by higher taxes may attract some to migrate to the region and cause others to leave. The net effect on earnings from the policy is all that is required to perform cost benefit calculations of the policy on outcomes.
- If the housing subsidy is the instrument, the issue of monotonicity is a red herring.
- If the subsidy is exogenously imposed, IV estimates the net effect of the policy on mean outcomes.
- Only if the effect of migration induced by the subsidy on outcomes is the question of interest, and not the effect of the subsidy, does uniformity emerge as an interesting question.

Comparing selection and IV models

- Angrist and Krueger (1999) compare IV with selection models and view the former with favor.
- Useful to understand this comparison in a model with essential heterogeneity.
- IV is estimating the derivative (or finite changes) of the parameters of a selection model.
- IV only conditions on Z (and X).

Comparing selection and IV models

- The control function approach conditions on Z and D (and X).
- From index sufficiency, equivalent to conditioning on $P(Z)$ and D:

$$\begin{aligned} & E(Y \mid X, D, Z) \\ & \quad = \mu_0(X) + [\mu_1(X) - \mu_0(X)] D \\ & \qquad\qquad + K_1(P(Z), X) D + K_0(P(Z), X)(1 - D) \end{aligned} \tag{30}$$

$$K_1(P(Z), X) = E(U_1 \mid D = 1, X, P(Z))$$

and

$$K_0(P(Z), X) = E(U_0 \mid D = 0, X, P(Z)).$$

Comparing selection and IV models

- IV approach does not condition on D.
- It works with the integral (over D) of (30).

$$
\begin{aligned}
& E\left(Y \mid X, P(Z)\right) \\
& \quad = \mu_0\left(X\right) + \left[\mu_1\left(X\right) - \mu_0\left(X\right)\right] P(Z) \\
& \qquad + K_1\left(P\left(Z\right), X\right) P(Z) + K_0\left(P\left(Z\right), X\right)\left(1 - P(Z)\right)
\end{aligned}
\tag{31}
$$

Under monotonicity and (A-1)–(A-5)

$$
\left.\frac{\partial E(Y \mid X, P(Z))}{\partial P(Z)}\right|_{P(Z)=p} = \mathrm{LIV}\left(X, p\right) = \mathrm{MTE}\left(X, p\right).
$$

- Control function builds up MTE from components.
- IV gets it in one fell swoop.

Comparing selection and IV models

- With rank and limit conditions (Heckman, 1990; Heckman and Robb, 1985), using control functions, one can identify $\mu_1(X)$, $\mu_0(X)$, $K_1(P(Z), X)$, and $K_0(P(Z), X)$.
- The selection (control function) estimator identifies the conditional means

$$E(Y_1 \mid X, P(Z), D = 1) = \mu_1(X) + K_1(X, P(Z)) \tag{32a}$$

and

$$E(Y_0 \mid X, P(Z), D = 0) = \mu_0(X) + K_0(X, P(Z)) . \tag{32b}$$

Comparing selection and IV models

- To decompose these means and separate $\mu_1(X)$ from $K_1(X, P(Z))$ without invoking functional form assumptions, it is necessary to have an exclusion (a Z not in X).
- This allows $\mu_1(X)$ and $K_1(X, P(Z))$ to be independently varied with respect to each other.
- We can also invoke curvature conditions without exclusion of variables.
- In addition there must exist a limit set for Z given X such that $K_1(X, P(Z)) = 0$ for Z in that limit set.

Comparing selection and IV models

- Limit set not required for selection model if we are interested only in MTE or LATE.
- Not required in IV either if we only seek MTE or LATE.

Comparing selection and IV models

- Without functional form assumptions, it is not possible to disentangle $\mu_1(X)$ from $K_1(X, P(Z))$ which may contain constants and functions of X that do not interact with $P(Z)$ (see Heckman (1990)).
- These limit set arguments are needed for ATE or TT, not LATE or LIV.

IV method

- IV method works with derivatives of (31) and not levels.
- Cannot directly recover the constant terms in (32a) and (32b).

IV method

- In summary, the control function method directly identifies levels while the LIV approach works with slopes.
- Constants that do not depend on $P(Z)$ disappear from the LIV estimates of the model.

IV method

- The distributions of U_1, U_0 and V do not need to be specified to estimate control function models (see Powell, 1994).
- In particular, there is no reliance on normality.

Support problems for IV

- Support conditions with control function models have their counterparts in IV models.
- One common criticism of selection models is that without invoking functional form assumptions, identification of $\mu_1(X)$ and $\mu_0(X)$ requires that $P(Z) \to 1$ and $P(Z) \to 0$ in limit sets.
- Identification in limit sets is sometimes called "identification at infinity."
- In order to identify $\text{ATE} = E(Y_1 - Y_0|X)$, IV methods also require that $P(Z) \to 1$ and $P(Z) \to 0$ in limit sets, so an identification at infinity argument is implicit when IV is used to identify this parameter.

Support problems for IV

- The LATE parameter avoids this problem by moving the goal posts and redefining the parameter of interest from a level parameter like ATE or TT to a slope parameter like LATE which differences out the unidentified constants.
- We can identify this parameter by selection models or IV models without invoking identification at infinity.

Support problems for IV

- The IV estimator is model dependent, just like the selection estimator, but in application, the model does not have to be fully specified to obtain Δ^{IV} using Z (or $J(Z)$).
- However the distribution of $P(Z)$ and the relationship between $P(Z)$ and $J(Z)$ generates the weights on MTE (or LIV).
- The interpretation placed on Δ^{IV} in terms of weights on Δ^{MTE} depends crucially on the specification of $P(Z)$. In both control function and IV approaches for the general model of heterogeneous responses, $P(Z)$ plays a central role.

Support problems for IV

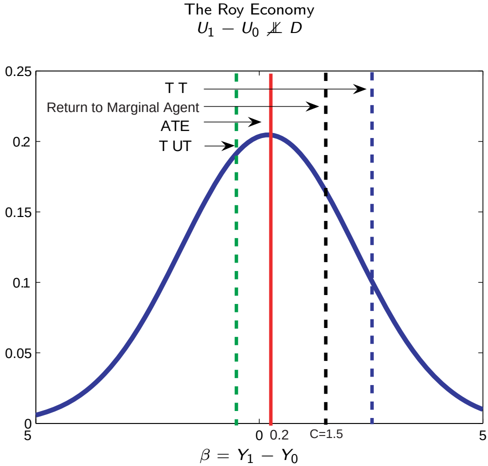
- Two economists using the same instrument will obtain the same point estimate using the same data.
- Their *interpretation* of that estimate will differ depending on how they specify the arguments in $P(Z)$, even if neither uses $P(Z)$ as an instrument.
- By conditioning on $P(Z)$, the control function approach makes the dependence of estimates on the specification of $P(Z)$ explicit.
- The IV approach is less explicit and masks the assumptions required to economically interpret the empirical output of an IV estimation.

Examples based on choice theory

- Suppose cost of adopting the policy C is the same across all countries.
- Countries choose to adopt the policy if $D^* > 0$ where D^* is the net benefit: $D^* = (Y_1 - Y_0 - C)$ and
- $ATE = E(\beta) = E(Y_1 - Y_0) = \mu_1 - \mu_0$
- Treatment on the treated is

$$\begin{aligned} E(\beta \mid D = 1) &= E(Y_1 - Y_0 \mid D = 1) \\ &= \mu_1 - \mu_0 + E(U_1 - U_0 \mid D = 1). \end{aligned}$$

Figure 1: distribution of gains



TT$= 2.666$, TUT$= -0.632$

Return to Marginal Agent $= C = 1.5$, ATE $= \mu_1 - \mu_0 = \bar{\beta} = 0.2$

The model

Outcomes	Choice Model
$Y_1 = \mu_1 + U_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \mu_0 + U_0 = \alpha + U_0$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$

General Case
$(U_1 - U_0) \not\perp\!\!\!\perp D$ $\text{ATE} \neq \text{TT} \neq \text{TUT}$

The model

The Researcher Observes (Y, D, C)

$$Y = \alpha + \beta D + U_0 \text{ where } \beta = Y_1 - Y_0$$

Parameterization

$$\begin{array}{lll} \alpha = 0.67 & (U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma}) & D^* = Y_1 - Y_0 - C \\ \bar{\beta} = 0.2 & \mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} & C = 1.5 \end{array}$$

- Let $C = \gamma Z$, $\gamma \geq 0$.

Figure 4A: monotonicity, the extended Roy economy
Standard case

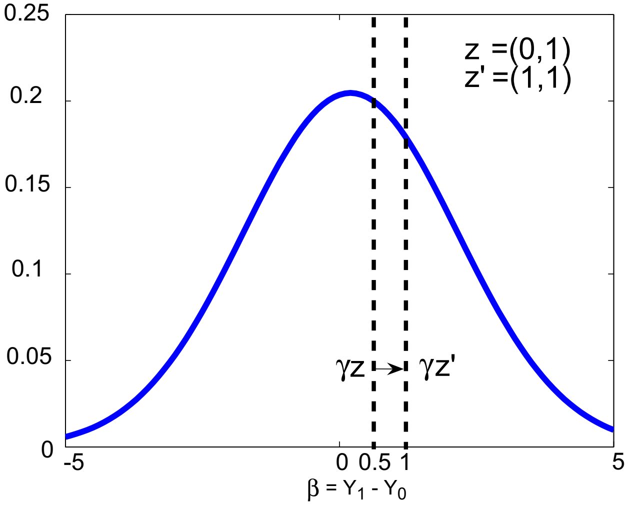


Figure 4B: monotonicity, the extended Roy economy
Changing Z_1 without controlling for Z_2

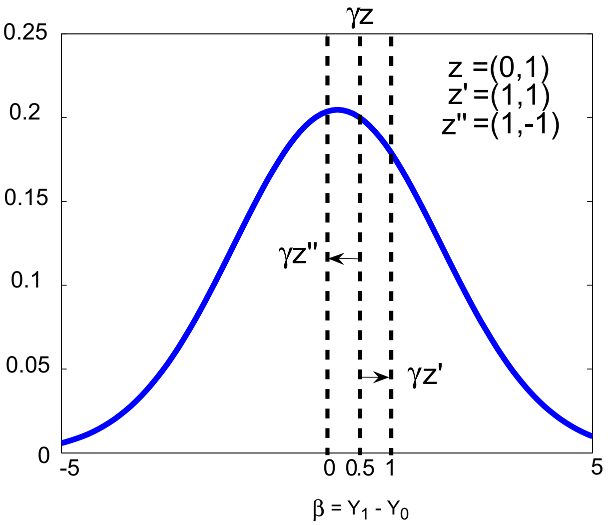


Figure 4C: monotonicity, the extended Roy economy
Random coefficient case

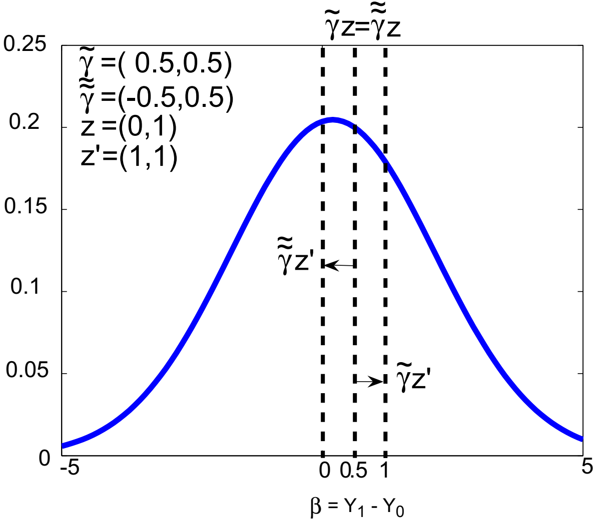


Figure 4: monotonicity, the extended Roy economy

A. Standard Case	B. Changing Z_1 without Controlling for Z_2	C. Random Coefficient Case
$z \longrightarrow z'$ $z = (0,1)$ and $z' = (1,1)$	$z \longrightarrow z'$ or $z \longrightarrow z''$ $z = (0,1)$, $z' = (1,1)$ and $z'' = (1,-1)$	$z \longrightarrow z'$ $z = (0,1)$ and $z' = (1,1)$
		γ is a random vector $\widetilde{\gamma} = (0.5, 0.5)$ and $\widetilde{\widetilde{\gamma}} = (-0.5, 0.5)$ where $\widetilde{\gamma}$ and $\widetilde{\widetilde{\gamma}}$ are two realizations of γ
$D(\gamma z) \geq D(\gamma z')$	$D(\gamma z) \geq D(\gamma z')$ or $D(\gamma z) < D(\gamma z'')$	$D\left(\widetilde{\widetilde{\gamma}} z\right) \geq D\left(\widetilde{\widetilde{\gamma}} z'\right)$ and $D\left(\widetilde{\gamma} z\right) < D\left(\widetilde{\gamma} z'\right)$
For all individuals	Depending on the value of z' or z''	Depending on value of γ

Figure 4: monotonicity, the extended Roy economy model

Outcomes	Choice Model
$Y_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \alpha + U_0$	$D = \begin{cases} 1 & \text{if} \quad Y_1 - Y_0 - \gamma Z > 0 \\ 0 & \text{if} \quad Y_1 - Y_0 - \gamma Z \leq 0 \end{cases}$ with $\gamma Z = \gamma_1 Z_1 + \gamma_2 Z_2$

Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma}) \ , \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \quad \begin{array}{l} \alpha = 0.67, \ \bar{\beta} = 0.2, \\ \gamma = (0.5, 0.5) \text{ (except in Case C)} \end{array}$$

$$Z_1 = \{-1, 0, 1\} \text{ and } Z_2 = \{-1, 0, 1\}$$

Figure 5: IV weights and its components under discrete instruments when $P(Z)$ is the instrument

$$
\begin{aligned}
&\Delta^{\text{LATE}}\left(p_\ell, p_{\ell+1}\right) \\
&\quad = \frac{E\left(Y|P(Z)=p_{\ell+1}\right) - E\left(Y|P(Z)=p_\ell\right)}{p_{\ell+1}-p_\ell} \\
&\quad = \frac{\overline{\beta}\left(p_{\ell+1}-p_\ell\right) + \sigma_{U_1-U_0}\left(\phi\left(\Phi^{-1}\left(1-p_{\ell+1}\right)\right) - \phi\left(\Phi^{-1}\left(1-p_\ell\right)\right)\right)}{p_{\ell+1}-p_\ell}
\end{aligned}
$$

$$
\begin{aligned}
\lambda_\ell &= \left(p_{\ell+1}-p_\ell\right) \frac{\sum_{i=1}^{K}\left(p_i - E\left(P\left(Z\right)\right)\right)\sum_{t>\ell}^{K} f\left(p_i, p_t\right)}{\operatorname{Cov}\left(Z_1, D\right)} \\
&= \left(p_{\ell+1}-p_\ell\right) \frac{\sum_{t>\ell}^{K}\left(p_t - E\left(P\left(Z\right)\right)\right) f\left(p_t\right)}{\operatorname{Cov}\left(Z_1, D\right)}
\end{aligned}
$$

Joint probability distribution of (Z_1, Z_2) and the propensity score

$Z_1 \backslash Z_2$	-1	0	1
-1	0.02	0.02	0.36
	0.7309	*0.6402*	*0.5409*
0	0.3	0.01	0.03
	0.6402	*0.5409*	*0.4388*
1	0.2	0.05	0.01
	0.5409	*0.4388*	*0.3408*

$$\text{Cov}(Z_1, Z_2) = -0.5468$$

(joint probabilities in ordinary type $(\Pr(Z_1 = z_1, Z_2 = z_2))$;
propensity score in italics $(\Pr(D = 1 | Z_1 = z_1, Z_2 = z_2)))$

Figure 5: IV weights and its components under discrete instruments when $P(Z)$ is the instrument

$$\text{ATE} = 0.2, \quad \text{TT} = 0.5942, \quad \text{TUT} = -0.4823$$

and

$$\Delta^{\text{IV}}_{P(Z)} = \sum_{\ell=1}^{K-1} \Delta^{\text{LATE}} \left(p_{\ell}, p_{\ell+1}\right) \lambda_{\ell} = -0.09$$

Figure 5A: IV weights and its components under discrete instruments when $P(Z)$ is the instrument (IV Weights)

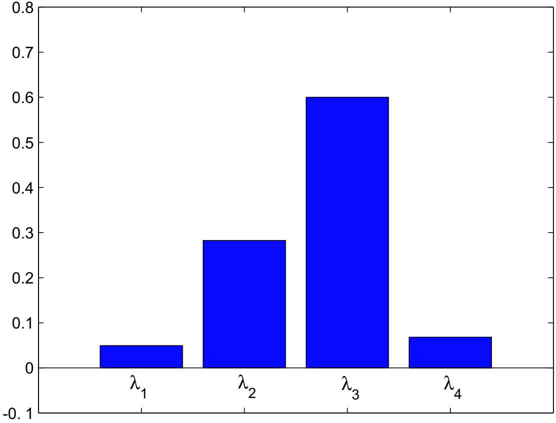


Figure 5B: IV weights and its components under discrete instruments when $P(Z)$ is the instrument $(E(P(Z) \mid P(Z) > p_\ell)$ and $E(P(Z)))$

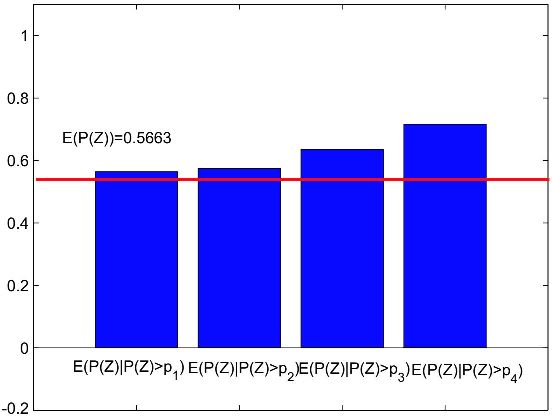
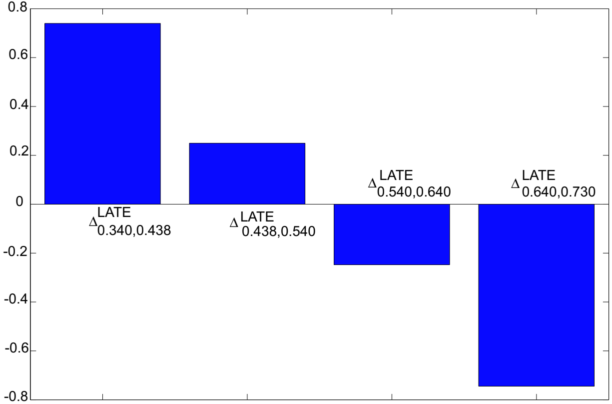


Figure 5C: IV weights and its components under discrete instruments when $P(Z)$ is the instrument (Local average treatment effects)



Consider using Z_1 as instrument

- If Z_1 and Z_2 are negatively dependent and $E(Z_1 \mid P(Z) > u_D)$ is not monotonic in u_D, weights negative.
- This nonmonotonicity is evident in Figure 6B.
- This produces the pattern of negative weights shown in Figure 6A.
- Associated with two way flows.
- Two way flows are induced by uncontrolled variation in Z_2.

Figure 4B: monotonicity, the extended Roy economy
Changing Z_1 without controlling for Z_2

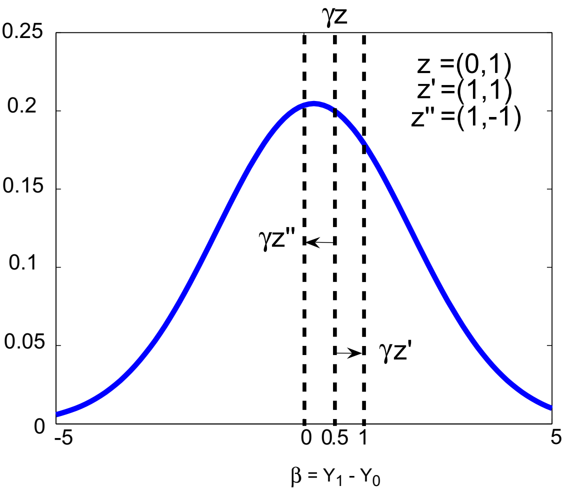
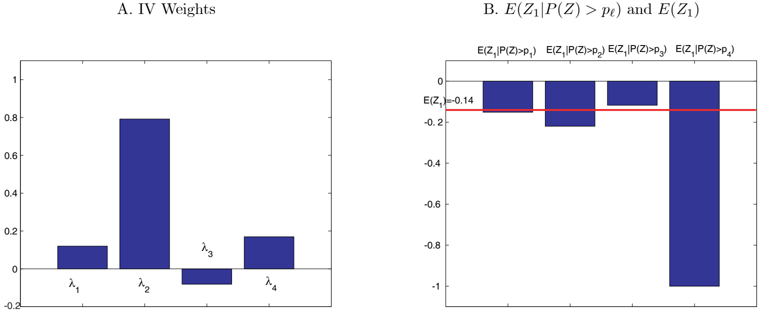
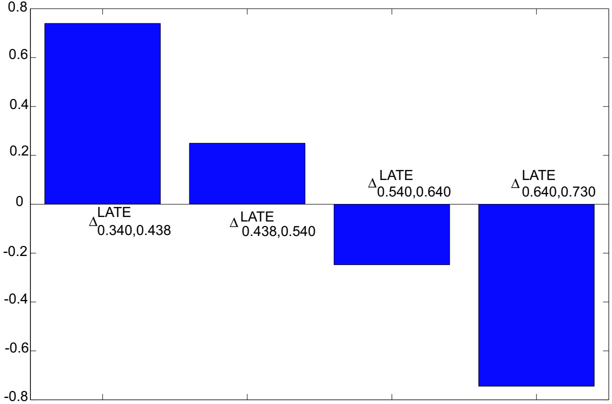


Figure 6: IV weights and its components under discrete instruments when Z_1 is the instrument



The model is the same as the one presented after figure 4.

Figure 5C: IV weights and its components under discrete instruments when $P(Z)$ is the instrument (local average treatment effects)



$$\Delta_{Z_1}^{\mathrm{IV}} = \sum_{\ell=1}^{K-1} \Delta^{\mathrm{LATE}} \left(p_\ell, p_{\ell+1}\right) \lambda_\ell = 0.1833$$

$$\lambda_\ell = \left(p_{\ell+1} - p_\ell\right) \frac{\sum_{i=1}^{I} \left(z_{1,i} - E\left(Z_1\right)\right) \sum_{t>\ell}^{K} f\left(z_{1,i}, p_t\right)}{\mathrm{Cov}\left(Z_1, D\right)}$$

Joint probability distribution of (Z_1, Z_2) and the propensity score

$Z_1 \backslash Z_2$	-1	0	1
-1	0.02	0.02	0.36
	0.7309	*0.6402*	*0.5409*
0	0.3	0.01	0.03
	0.6402	*0.5409*	*0.4388*
1	0.2	0.05	0.01
	0.5409	*0.4388*	*0.3408*

$$\mathrm{Cov}(Z_1, Z_2) = -0.5468$$

(joint probabilities in ordinary type $(\Pr(Z_1 = z_1, Z_2 = z_2))$;
propensity score in italics $(\Pr(D = 1 | Z_1 = z_1, Z_2 = z_2)))$

Conditional variable estimator and conditional local average treatment effect when Z_1 is the instrument (given $Z_2 = z_2$)

	$Z_2 = -1$	$Z_2 = 0$	$Z_2 = 1$
$P(-1, Z_2) = p_3$	0.7309	0.6402	0.5409
$P(0, Z_2) = p_2$	0.6402	0.5409	0.4388
$P(1, Z_2) = p_1$	0.5409	0.4388	0.3408
λ_1	0.8418	0.5384	0.2860
λ_2	0.1582	0.4616	0.7140
$\Delta^{\text{LATE}}(p_1, p_2)$	-0.2475	0.2497	0.7470
$\Delta^{\text{LATE}}(p_2, p_3)$	-0.7448	-0.2475	0.2497
$\Delta^{\text{IV}}_{Z_1 \mid Z_2 = z_2}$	-0.3262	0.0202	0.3920

Conditional instrumental variable estimator

$$\Delta^{\mathrm{IV}}_{Z_1|Z_2=z_2} = \sum_{\ell=1}^{I-1} \Delta^{\mathrm{LATE}}\left(p_\ell, p_{\ell+1}|Z_2 = z_2\right) \lambda_{\ell|Z_2=z_2} = \sum_{\ell=1}^{I-1} \Delta^{\mathrm{LATE}}\left(p_\ell, p_{\ell+1}|Z_2 = z_2\right) \lambda_{\ell|Z_2=z_2}$$

$$\Delta^{\mathrm{LATE}}\left(p_\ell, p_{\ell+1}|Z_2 = z_2\right) = \frac{E\left(Y|P(Z) = p_{\ell+1}, Z_2 = z_2\right) - E\left(Y|P(Z) = p_\ell, Z_2 = z_2\right)}{p_{\ell+1} - p_\ell}$$

$$\begin{aligned}
\lambda_{\ell|Z_2=z_2} &= \left(p_{\ell+1} - p_\ell\right) \frac{\sum_{i=1}^{I} \left(z_{1,i} - E\left(Z_1|Z_2 = z_2\right)\right) \sum_{t>\ell}^{I} f\left(z_{1,i}, p_t|Z_2 = z_2\right)}{\mathrm{Cov}\left(Z_1, D\right)} \\
&= \left(p_{\ell+1} - p_\ell\right) \frac{\sum_{t>\ell}^{I} \left(z_{1,t} - E\left(Z_1|Z_2 = z_2\right)\right) f\left(z_{1,t}, p_t|Z_2 = z_2\right)}{\mathrm{Cov}\left(Z_1, D\right)}
\end{aligned}$$

Conditional instrumental variable estimator

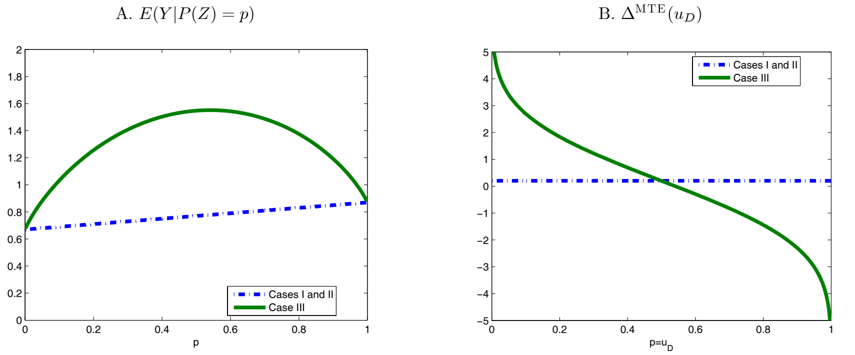
Probability Distribution of Z_1 Conditional on Z_2 $(\Pr(Z_1 = z_1|Z_2 = z_2))$

z_1	$\Pr(Z_1 = z_1\|Z_2 = -1)$	$\Pr(Z_1 = z_1\|Z_2 = 0)$	$\Pr(Z_1 = z_1\|Z_2 = 1)$
-1	0.0385	0.25	0.9
0	0.5769	0.125	0.075
1	0.3846	0.625	0.025

Continuous instruments

- Figure 7 plots $E(Y \mid P(Z))$ and MTE for the models displayed at the base of the figure. In cases I and II, $\beta \perp\!\!\!\perp D$.
- In case I, this is trivial since β is a constant. In case II, β is random but selection into D does not depend on β.
- Case III is the model with essential heterogeneity ($\beta \not\perp\!\!\!\perp D$).
- Figure 7A depicts $E(Y \mid P(Z))$ in the three cases.

Figure 7: conditional expectation of Y on $P(Z)$ and the marginal treatment effect (MTE)



Outcomes	Choice Model
$Y_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \alpha + U_0$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$

Case I	Case II	Case III
$U_1 = U_0$ $\bar{\beta}$ =ATE=TT=TUT=IV	$U_1 - U_0 \perp\!\!\!\perp D$ $\bar{\beta}$ =ATE=TT=TUT=IV	$U_1 - U_0 \not\perp\!\!\!\perp D$ $\bar{\beta}$ =ATE≠TT≠TUT≠IV

Parameterization

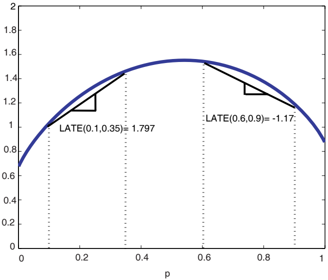
Cases I, II and III	Cases II and III	Case III
$\alpha = 0.67$ $\bar{\beta} = 0.2$	$(U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$	$D^* = Y_1 - Y_0 - \gamma Z$ $Z \sim N(\mu_Z, \mathbf{\Sigma}_Z)$ $\mu_Z = (2, -2)$ and $\mathbf{\Sigma}_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$ $\gamma = (0.5, 0.5)$

- Cases I and II make $E(Y \mid P(Z))$ linear in $P(Z)$ (see equation 23). Case III is nonlinear in $P(Z)$ which arises when $\beta \not\perp\!\!\!\perp D$. The derivative of $E(Y \mid P(Z))$ is presented in the right panel (Figure 7B).
- It is a constant in cases I and II (flat MTE) but declining in $U_D = P(Z)$ for the case with selection on the gain.

- MTE gives the mean marginal return for persons who have utility $P(Z) = u_D$ ($P(Z) = u_D$ is the margin of indifference).
- Figure 7 highlights that MTE (and LATE) identify average returns for persons at the margin of indifference at different levels of the mean utility function $P(Z)$.
- Figure 8 plots MTE and LATE for different intervals of u_D using the model plotted in Figure 7.
- LATE is the chord of $E(Y \mid P(Z))$ evaluated at different points.
- The relationship between LATE and MTE is presented in the right panel of Figure 8.

Figure 8: the local average treatment effect

A. $E(Y|P(Z)=p)$ and $\Delta^{\text{LATE}}(p_\ell, p_{\ell+1})$



B. $\Delta^{\text{MTE}}(u_D)$ and $\Delta^{\text{LATE}}(p_\ell, p_{\ell+1})$

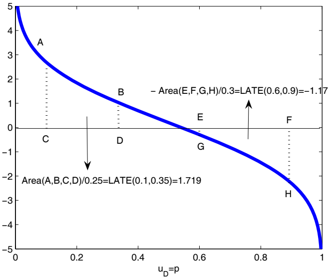


Figure 8: the local average treatment effect

$$\Delta^{\text{LATE}}(p_\ell, p_{\ell+1}) = \frac{E\left(Y|P(Z)=p_{\ell+1}\right) - E\left(Y|P(Z)=p_\ell\right)}{p_{\ell+1}-p_\ell}$$

$$= \frac{\int_{p_\ell}^{p_{\ell+1}} \Delta^{\text{MTE}}(u_D)\,du_D}{p_{\ell+1}-p_\ell}$$

$$\begin{aligned} \Delta^{\text{LATE}}(0.1, 0.35) &= 1.719 \\ \Delta^{\text{LATE}}(0.6, 0.9) &= -1.17 \end{aligned}$$

Figure 8: the local average treatment effect

Outcomes	Choice Model
$Y_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \alpha + U_0$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$ with $D^* = Y_1 - Y_0 - \gamma Z$

Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma}) \text{ and } Z \sim N(\mu_Z, \mathbf{\Sigma}_Z)$$

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \ \mu_Z = (2, -2) \text{ and } \mathbf{\Sigma}_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$$

$$\alpha = 0.67, \bar{\beta} = 0.2, \gamma = (0.5, 0.5)$$

- The treatment parameters as a function of p associated with case III are plotted in Figure 9.
- MTE is the same as that reported in Figure 7.
- ATE is the same for all p.
- $\Delta^{TT}(p) = E(Y_1 - Y_0 \mid D = 1, P(Z) = p)$ declines in p (equivalently, it declines in u_D).

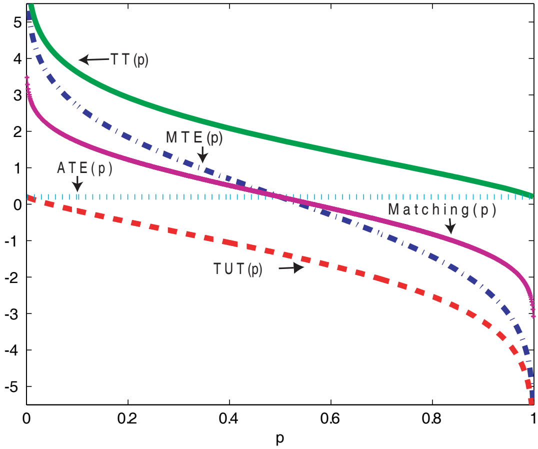
$$\begin{aligned} LATE(p, p') &= \frac{\Delta^{TT}(p')p' - \Delta^{TT}(p)p}{p' - p}, \qquad p' \neq p \\ MTE &= \frac{\partial[\Delta^{TT}(p)p]}{\partial p}. \end{aligned}$$

Parameter	Definition	Under Assumptions (*)
Marginal Treatment Effect	$E\left[Y_1 - Y_0 \mid D^* = 0, P(Z) = p\right]$	$\bar{\beta} + \sigma_{U_1 - U_0}\Phi^{-1}(1-p)$
Average Treatment Effect	$E\left[Y_1 - Y_0 \mid P(Z) = p\right]$	$\overline{\beta}$
Treatment on the Treated	$E\left[Y_1 - Y_0 \mid D^* > 0, P(Z) = p\right]$	$\bar{\beta} + \sigma_{U_1 - U_0}\frac{\phi\left(\Phi^{-1}(1-p)\right)}{p}$
Treatment on the Untreated	$E\left[Y_1 - Y_0 \mid D^* \leq 0, P(Z) = p\right]$	$\bar{\beta} - \sigma_{U_1 - U_0}\frac{\phi\left(\Phi^{-1}(1-p)\right)}{1-p}$
OLS/Matching on $P(Z)$	$E\left[Y_1 \mid D^* > 0, P(Z) = p\right] - E\left[Y_0 \mid D^* \leq 0, P(Z) = p\right]$	$\bar{\beta} + \left(\frac{\sigma^2_{U_1} - \sigma_{U_1,U_0}}{\sqrt{\sigma_{U_1-U_0}}}\right)\left(\frac{1-2p}{p(1-p)}\right)\phi\left(\Phi^{-1}(1-p)\right)$

Note: $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively. $\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$.

(*): The model in this case is the same as the one presented below Figure 6.

Figure 9: treatment parameters and OLS matching as a function of $P(Z) = p$



Another nonmonotonicity example

A mixture of two normals:

$$Z \sim P_1 N(\mu_1, \Sigma_1) + P_2 N(\mu_2, \Sigma_2)$$

P_1 is the proportion in population 1, P_2 is the proportion in population 2 and $P_1 + P_2 = 1$.

Another nonmonotonicity example

- Conventional normal outcome selection model generated by the parameters at the base of Figure 11.
- The discrete choice equation is a conventional probit: $\Pr(D = 1 \mid Z = z) = \Phi\left(\frac{\gamma z}{\sigma_V}\right)$.
- The $\Delta^{\text{MTE}}(v)$,

$$E(Y_1 - Y_0 \mid V = v) = \mu_1 - \mu_0 + \frac{Cov(U_1 - U_0, V)}{Var(V)} v.$$

- We show results for models with vector Z that satisfies (IV-1) and (IV-2) and with $\gamma > 0$ componentwise.

Outcomes	Choice Model
$Y_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \alpha + U_0$	$D = \begin{cases} 1 & \text{if} \quad D^* > 0 \\ 0 & \text{if} \quad D^* \leq 0 \end{cases}$ $D^* = Y_1 - Y_0 - \gamma Z$ and $V = -(U_1 - U_0)$

Parameterization

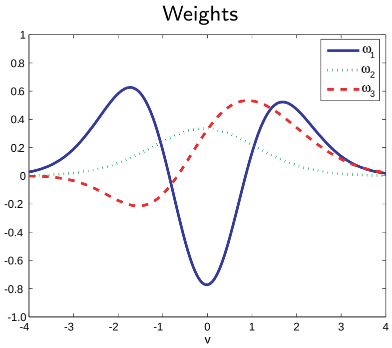
$$(U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \quad \alpha = 0.67, \ \bar{\beta} = 0.2$$

$$Z = (Z_1, Z_2) \sim p_1 N(\kappa_1, \Sigma_1) + p_2 N(\kappa_2, \Sigma_2)$$

$$p_1 = 0.45,\ \ p_2 = 0.55 \qquad ; \quad \Sigma_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.4 \end{bmatrix}$$

$$\mathrm{Cov}(Z_1, \gamma Z) = \gamma \Sigma_1^1 = 0.98 \quad ; \qquad \gamma = (0.2, 1.4)$$

Figure 11: marginal treatment effect and IV weights using Z_1 as the instrument when $Z = (Z_1, Z_2) \sim p_1 N(\mu_1, \Sigma_1) + p_2 N(\mu_2, \Sigma_2)$ for different values of Σ_2



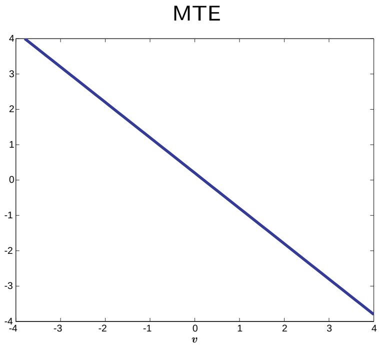


Table 3: IV estimator and $\mathrm{Cov}(Z_2, \gamma' Z)$ associated with each value of Σ_2

Weights	Σ_2	κ_1	κ_2	IV	ATE	TT	TUT	$\mathrm{Cov}(Z_2, \gamma Z) = \gamma \Sigma_2^1$
ω_1	$\begin{bmatrix} 0.6 & -0.5 \\ -0.5 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	0.434	0.2	1.401	−1.175	−0.58
ω_2	$\begin{bmatrix} 0.6 & 0.1 \\ 0.1 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	0.078	0.2	1.378	−1.145	0.26
ω_3	$\begin{bmatrix} 0.6 & -0.3 \\ -0.3 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \end{bmatrix}$	−2.261	0.2	1.310	−0.859	−0.30

Consider the study of the GED.

Figure 12: frequency of the propensity score by final schooling decision

Dropouts and GEDs – males of the NLSY at age 30

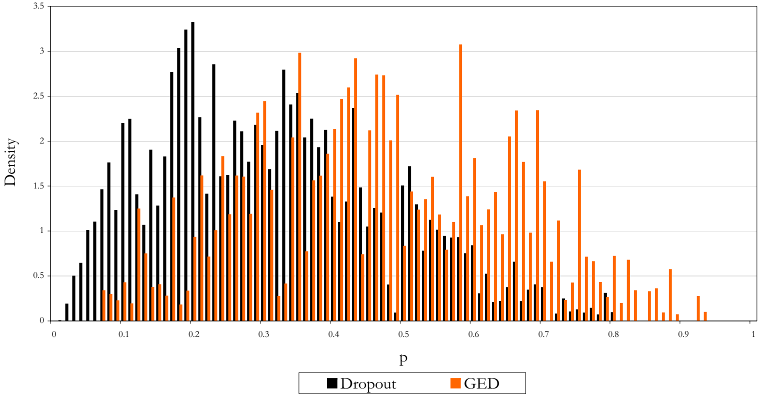
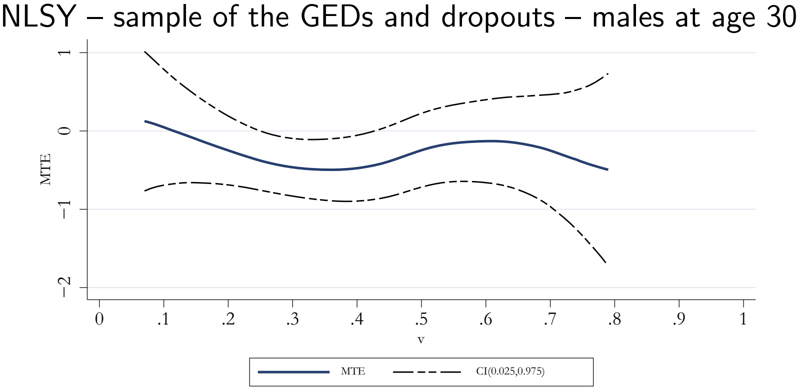


Table 4: instrumental variables estimates

Sample of GEDs and dropouts – males at age 30

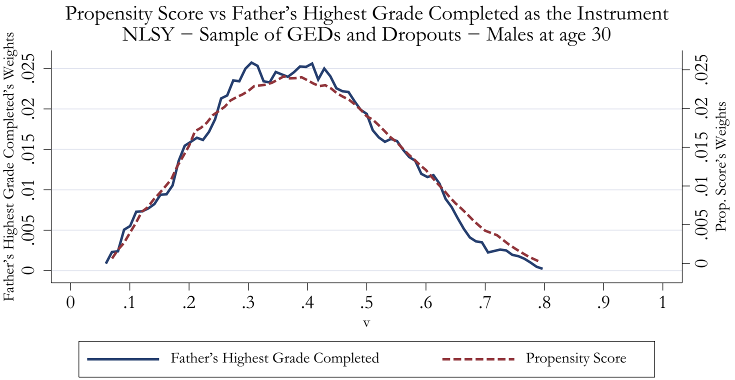
Instruments	Standard IV [(f)]	
	Full Sample[(a)]	Common Support[(b)]
Father Highest Grade Completed	0.194	0.005
	(0.384)	*(0.391)*
Mother Highest Grade Completed	1.106	0.588
	(3.030)	*(2.981)*
Number of Siblings	-0.311	-0.471
	(0.618)	*(0.725)*
Ged Cost	1.938	1.994
	(2.414)	*(2.544)*
Family income in 1979	0.656	0.636
	(0.534)	*(0.571)*
Dropout's local wage at age 17	-1.812	-1.612
	(1.228)	*(1.037)*
High School Graduate's local wage at age 17	-2.197	-1.872
	(1.441)	*(1.143)*
Dropout's local unemployment rate at age 17	0.164	0.203
	(1.071)	*(0.853)*
High School Graduate's local unemployment rate at age 17	0.142	0.202
	(1.537)	*(1.261)*
Propensity Score [(d)]	-0.276	-0.305
	(0.134)	*(0.140)*

Figure 13: MTE of the GED with confidence interval



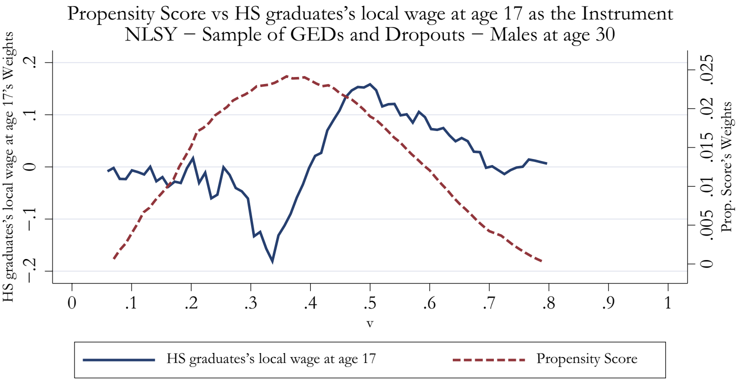
The dependent variable in the outcome equation is hourly earnings at age 30. The controls in the outcome equations are tenure, tenure squared, experience, corrected AFQT, black (dummy), hispanic (dummy), marital status, and years of schooling. Let D=0 denote dropout status, and D=1 denote GED status. The model for D (choice model) includes as controls the corrected AFQT, number of siblings, father's education, mother's education, family income at age 17, local GED costs, broken home at age 14, average local wage at age 17 for dropouts and high school graduates, local unemployment rate at age 17 for dropouts and high school graduates, the dummy variables black and hispanics, and a set of dummy variables controlling for the year of birth. The choice model is estimated using a probit model. In computing the MTE, the bandwidth in the first step is selected using the leave−one−out cross−validation method. In the second step, following Carneiro (2003) and Heckman et.al. (1998), we set the bandwidht to 0.3. We use biweight kernel functions.

Figure 14: IV weights



The dependent variable in the outcome equation is hourly earnings at age 30. The controls in the outcome equations are tenure, tenure squared, experience, corrected AFQT, black (dummy), hispanic (dummy), marital status, and years of schooling. Let D=0 denote dropout status, and D=1 denote GED status. The model for D (choice model) includes as controls the corrected AFQT, number of siblings, father's education, mother's education, family income at age 17, local GED costs, broken home at age 14, average local wage at age 17 for dropouts and high school graduates, local unemployment rate at age 17 for dropouts and high school graduates, the dummy variables black and hispanics, and a set of dummy variables controlling for the year of birth. The choice model is estimated using a probit model. In computing the MTE, the bandwidth in the first step is selected using the leave−one−out cross−validation method. In the second step, following Carneiro (2003) and Heckman et.al. (1998), we set the bandwidht to 0.3. We use biweight kernel functions.

Figure 15: IV weights



The dependent variable in the outcome equation is hourly earnings at age 30. The controls in the outcome equations are tenure, tenure squared, experience, corrected AFQT, black (dummy), hispanic (dummy), marital status, and years of schooling. Let D=0 denote dropout status, and D=1 denote GED status. The model for D (choice model) includes as controls the corrected AFQT, number of siblings, father's education, mother's education, family income at age 17, local GED costs, broken home at age 14, average local wage at age 17 for dropouts and high school graduates, local unemployment rate at age 17 for dropouts and high school graduates, the dummy variables black and hispanics, and a set of dummy variables controlling for the year of birth. The choice model is estimated using a probit model. In computing the MTE, the bandwidth in the first step is selected using the leave−one−out cross−validation method. In the second step, following Carneiro (2003) and Heckman et.al. (1998), we set the bandwidht to 0.3. We use biweight kernel functions.

Table 5: treatment parameter estimates

Sample of GED and Dropouts - Males at age 30 [(a)]

Treatment Parameter	Parametric [(b)]	Polynomial [(c)]	Nonparametric [(d)]
Treatment on the Treated	-0.152	-0.183	-0.241
	(0.166)	*(0.201)*	*(0.180)*
Treatment on the Untreated	-0.369	-0.119	-0.304
	(0.170)	*(0.231)*	*(0.223)*
Average Treatment Effect	-0.279	-0.145	-0.278
	(0.151)	*(0.184)*	*(0.174)*
LATE(0.38,0.62)	-0.335	-0.404	-0.261
	(0.160)	*(0.275)*	*(0.221)*
LATE(0.55,0.79)	-0.453	0.106	-0.327
	(0.205)	*(0.377)*	*(0.416)*
LATE(0.21,0.45)	-0.216	-0.462	-0.396
	(0.153)	*(0.210)*	*(0.164)*

Notes: (a) We excluded the oversample of poor whites, the military sample, and those who attended college. (b) The treatment parameters are estimated by taking the weighted sum of the MTE estimated using the parametric approach. (c) The treatment parameters are estimated by taking the weighted sum of the MTE estimated using A polynomial of degree 4 to approximate $E(Y|P)$. (d) The treatment parameters are estimated by taking the weighted sum of the MTE estimated using the nonparametric approach. The standard deviations (in parenthesis) are computed using bootstrapping (100 draws).

Relaxing additive separability in the choice equation and allowing for random coefficient choice models

- The analysis of this lecture and the entire recent literature on instrumental variables estimators for models with essential heterogeneity relies on the assumption that the treatment choice equation is in additively separable form (14).
- Imparts an asymmetry to the entire instrumental variable enterprise for estimating treatment effects.

Relaxing additive separability in the choice equation and allowing for random coefficient choice models

- This asymmetry is also present in conventional selection models even in their semiparametric version.
- Parameters can be defined as weighted averages of an MTE but MTE and the derived parameters cannot be identified using any instrumental variables strategy.

Relaxing additive separability in the choice equation and allowing for random coefficient choice models

- Natural benchmark nonseparable model:
  - random coefficient model of choice $D = \mathbf{1}(\gamma Z \geq 0)$
  - γ is a random coefficient vector and $\gamma \perp\!\!\!\perp (Z, U_0, U_1)$.

Relaxing additive separability in the choice equation and allowing for random coefficient choice models

- Consider a more general case.
- Relax the separability assumption of equation (14).

$$D^* = \mu_D(Z, V), \quad D = \mathbf{1}(D^* \geq 0), \tag{33}$$

$\mu_D(Z, V)$ is not necessarily additively separable in Z and V, and V is not necessarily a scalar.

Relaxing additive separability in the choice equation and allowing for random coefficient choice models

We maintain assumptions (A-1)–(A-2) and (A-5).

- As we have shown, relationships among treatment parameters as weighted averages of generator functions (not MTEs) hold in this case even if we fail monotonicity.

Figure 4C: monotonicity, the extended Roy economy Random coefficient case

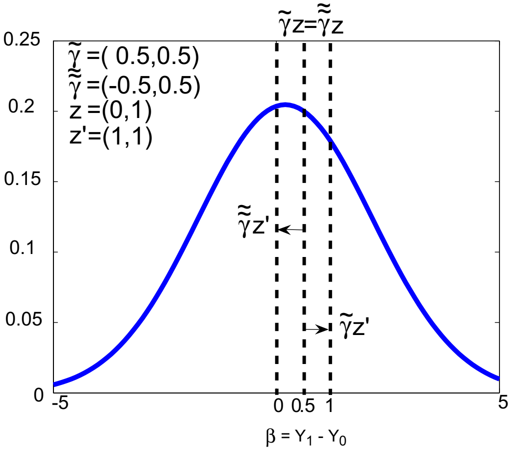


Figure 4C: monotonicity, the extended Roy economy Random coefficient case

$$z \longrightarrow z'$$
$$z = (0,1) \text{ and } z' = (1,1)$$

$$\gamma \text{ is a random vector}$$
$$\widetilde{\gamma} = (0.5, 0.5) \text{ and } \widetilde{\widetilde{\gamma}} = (-0.5, 0.5)$$
$$\text{where } \widetilde{\gamma} \text{ and } \widetilde{\widetilde{\gamma}} \text{ are two realizations of } \gamma$$

$$D\left(\widetilde{\widetilde{\gamma}}z\right) \geq D\left(\widetilde{\widetilde{\gamma}}z'\right) \text{ and } D\left(\widetilde{\gamma}z\right) < D\left(\widetilde{\gamma}z'\right)$$

Depending on value of γ

Relaxing additive separability in the choice equation and allowing for random coefficient choice models

- In the additively separable case, MTE has three equivalent interpretations:
  1. $U_D (= F_V(V))$ is the only unobservable in the first stage decision rule, and MTE is the average effect of treatment given the unobserved characteristics in the decision rule $(U_D = u_D)$;
  2. MTE is the average effect of treatment given that the individual would be indifferent between treatment or not if $P(Z) = u_D$, where $P(Z)$ is a mean utility function;
  3. the MTE is an average effect conditional on the additive error term from the first stage choice model.

Relaxing additive separability in the choice equation and allowing for random coefficient choice models

- Under all interpretations of the MTE, and under the assumptions (A-1)–(A-5), MTE can be identified by LIV.
- Three definitions are not the same in the general nonseparable case (33). Heckman and Vytlacil (2001, 2005) extend MTE to the nonseparable case.

Failure of index sufficiency in general nonseparable models

- For any version of the nonseparable model, index sufficiency fails.
- Define $\Omega(z) = \{v : \mu_D(z, v) \geq 0\}$.
- In the additively separable case, $P(z) \equiv \Pr(D = 1 \mid Z = z)$ $= \Pr(V_D \in \Omega(z))$, $P(z) = P(z') \Leftrightarrow \Omega(z) = \Omega(z')$.

- This produces index sufficiency so the propensity score orders the unobservables generating choices.
- In the more general case (33), it is possible to have (z, z') values such that $P(z) = P(z')$ and $\Omega(z) \neq \Omega(z')$ so index sufficiency does not hold.
- The Z's enter the model more generally, and the propensity score no longer plays the central role it plays in separable models.

The support of the propensity score

- The nonseparable model can also restrict the support of $P(Z)$.
- For example, consider a normal random coefficient choice model with a scalar regressor ($Z = (1, Z_1)$).
- Assume $\gamma_0 \sim N(0, \sigma_0^2)$, $\gamma_1 \sim N(\bar{\gamma}_1, \sigma_1^2)$, and $\gamma_0 \perp\!\!\!\perp \gamma_1$.

$$P(z_1) = \Phi\left(\frac{\bar{\gamma}_1 z_1}{\sqrt{\sigma_0^2 + \sigma_1^2 z_1^2}}\right).$$

- Φ is the cumulative distribution of a standard normal.
- $\sigma_1^2 > 0$.

- The support is strictly within the unit interval.
- The case when $\sigma_0^2 = 0$, the support is one point,

$$\left(P\left(z\right) = \Phi\left(\frac{\bar{\gamma}_1}{\sigma_1}\right)\right).$$

- Cannot, in general, identify ATE, TT or any treatment effect requiring the endpoints 0 or 1 using IV or control function strategies.

Violations of uniformity

- One source of violations of monotonicity is nonseparability between Z and V in (33).
- The random coefficient model is one intuitive model where separability fails.
- Even if (33) is separable in Z and V, uniformity may fail in the case of vector Z, where we use only one function of Z as the instrument, and do not condition on the remaining sources of variation in Z.
- If we condition appropriately, we retain monotonicity but get a new form of instrumental variable estimator that is sensitive to the specification of the Z not used as an instrument.

Summary and conclusion

- We have studied the estimation of treatment effects in a model

$$Y = \alpha + \beta D + \varepsilon$$

- We have contrasted this with a structural Roy model.
- Considered cases where β is constant and where β is heterogeneous.
- In the heterogeneous case $D \not\perp\!\!\!\perp \varepsilon$; $\beta \not\perp\!\!\!\perp D$; $\beta \not\perp\!\!\!\perp \varepsilon$.

Summary and conclusion

- Consider what IV estimates and its relationship with Economic Choice and Selection Models.
- In general heterogeneous response models, the two approaches have strong similarities.
- Selection models identify levels (conditional means).
- IV models identify slopes.

Summary and conclusion

- We lose constants in estimating IV models.
- We get back level parameters by integration.
- This accounts for the weighting schemes that appear in the literature.
- We must recover the constants to get levels parameters. (Classical treatment effects like ATE and TT).
- We restore the constants to estimate classical treatment parameters using the same limit arguments used to identify selection models.

Summary and conclusion

- If we are only concerned with slope treatment parameters, we can avoid limit arguments in IV or selection models.
- Explore the role of "monotonicity" or "uniformity" assumptions in IV.
- Concept used by Imbens and Angrist (1994) to define LATE.
- Monotonicity is not needed to define treatment parameters or establish the relationship among them (Heckman and Vytlacil).
- Under monotonicity or uniformity, LIV $=$ MTE.

Summary and conclusion

- Can express all classical treatment parameters as weighted averages of MTE.
- Monotonicity is needed to use IV to identify MTE and LATE.
- Treatment parameters can be defined; relationships among them established and IV weights defined without monotonicity or uniformity.

Summary and conclusion

- Much of the literature is for two outcome models.
- Angrist and Imbens (1995) consider the case of an ordered choice model with a scalar instrument that affects choices at all margins.
- We develop the case of a general ordered choice model with transition-specific instruments.
- We also develop a general unordered model.
- The most general case requires a marriage of semiparametric selection models (e.g. Heckman, 1990) and IV intuition to identify general parameters.

Summary and conclusion

- Need to identify semiparametric discrete choice models to get classical pairwise properties.
- We have an analysis for bounds which we defer to another occasion.

Extensions to More than Two Outcomes

- Angrist and Imbens (1995) extend their analysis of LATE to an ordered choice model with outcomes generated by a scalar instrument that can assume multiple values.
- From their analysis of the effect of schooling on earnings, it is unclear even under a strengthened "monotonicity" condition, whether IV estimates the effect of a change of schooling on earnings for a well defined margin of choice.
- To summarize their analysis, let $\bar{S}$ be the number of possible outcome states with associated outcomes Y_s and choice indicators D_s, $s = 1, \ldots, \bar{S}$.

- The s in their analysis correspond to different levels of schooling.
- For any two instrument values $Z = z_i$ and $Z = z_j$ with $z_i > z_j$, we can define associated indicators $\{D_s(z_i)\}_{s=1}^{\bar{S}}$ and $\{D_s(z_j)\}_{s=1}^{\bar{S}}$, where $D_s(z_i) = 1$ if a person assigned instrument value z_i chooses state s.
- As in the two outcome model, the instrument Z is assumed to be independent of the potential outcomes $\{Y_s\}_{s=1}^{\bar{S}}$ as well as the associated indicator functions defined by fixing Z at z_i and z_j.
- Observed schooling for instrument z_j is $S(z_j) = \sum_{s=1}^{\bar{S}} sD_s(z_j)$.
- Observed outcomes with this instrument are $Y(z_j) = \sum_{s=1}^{\bar{S}} Y_s D_s(z_j)$.

- Angrist and Imbens show that IV (with $Z = z_i$ and z_j) applied to S in a two stage least squares regression of Y on S identifies a "causal parameter"

$$\Delta^{\mathrm{IV}} = \sum_{s=2}^{\bar{S}} \left\{ E\left(Y_s - Y_{s-1} \mid S(z_i) \geq s > S(z_j)\right)\right\} \times \frac{\Pr\left(S(z_i) \geq s > S(z_j)\right)}{\sum_{s=2}^{\bar{S}} \Pr\left(S(z_i) \geq s > S(z_j)\right)}. \quad (34)$$

- This "causal parameter" is a weighted average of the gross return from going from $s-1$ to s for persons induced by the change in the instrument to move from *any* schooling level below s to *any* schooling level s or above.

- Thus the conditioning set defining the s component of IV includes people who have schooling below $s-1$ at instrument value $Z=z_j$ and people who have schooling above level s at instrument value $Z=z_i$.
- In this sum, the average return experienced by some of the people in the conditioning set for each component conditional expectation does not correspond to the average outcome corresponding to the gain in the argument of the expectation.
- In the case where $\bar{S}=2$, agents face only two choices and the margin of choice is well defined.

- Agents in each conditioning set are at different margins of choice.
- The weights are positive but, as noted by Angrist and Imbens, persons can be counted multiple times in forming the weights.
- When they generalize their analysis to multiple-valued instruments, they use the Yitzhaki (1989) weights.

- Whereas the weights in equation (34) can be constructed empirically, the terms in braces cannot be identified by any standard IV procedure.
- We present decompositions with components that are recoverable, whose weights can be estimated from the data and that are economically interpretable.

- We generalize LATE to a multiple outcome case where we can identify agents at different well defined margins of choice.
- Specifically, we (1) analyze both ordered and unordered choice models; (2) analyze outcomes associated with choices at various well defined margins; and (3) develop models with multiple instruments that can affect different margins of choice differently.
- With our methods, we can define and estimate a variety of economically interpretable parameters whereas the Angrist-Imbens analysis produces a single "causal parameter" (34) that does not answer any well defined policy problem.
- We first consider an explicit ordered choice model and decompose the IV into policy useful, identifiable, components.

Analysis of an Ordered Choice Model

- Ordered choice models arise in many settings.
- In schooling models, there are multiple grades.
- One has to complete grade $s - 1$ to proceed to grade s.
- The ordered choice model has been widely used to fit data on schooling transitions (Cameron and Heckman, 1998; Harmon and Walker, 1999).

- Its nonparametric identifiability has been studied (Carneiro, Hansen, and Heckman, 2003) and Cunha and Heckman (2007).
- It can also be used as a duration model for dynamic treatment effects with associated outcomes as in Cunha and Heckman (2007).
- It also represents the "vertical" model of the choice of product quality (Bresnahan, 1987; Prescott and Visscher, 1977; Shaked and Sutton, 1982) .

- Our analysis generalizes the preceding analysis for the binary model in a parallel way.
- Write potential outcomes as

$$Y_s = \mu_s(X, U_s) \qquad s = 1, \dots, \bar{S}.$$

The $\bar{S}$ could be different schooling levels or product qualities.

- We define latent variables $D_S^* = \mu_D(Z) - V$ where

$$D_s = \mathbf{1}[C_{s-1}(W_{s-1}) < \mu_D(Z) - V \leq C_s(W_s)], \qquad s = 1, \dots, \bar{S},$$

and the cutoff values satisfy

$$C_{s-1}(W_{s-1}) \leq C_s(W_s), \quad C_0(W_0) = -\infty \quad \text{and} \quad C_{\bar{S}}(W_{\bar{S}}) = \infty.$$

- The cutoffs used to define the intervals are allowed to depend on observed (by the economist) regressors W_s.

- We extend the analysis to allow the cutoffs to depend on unobserved regressors as well, following structural analysis along these lines by Carneiro et al. (2003) and Cunha and Heckman (2007). Observed outcomes are: $Y = \sum_{s=1}^{\bar{S}} Y_s D_s$.
- The Z shift the index generally, the W_s affect s-specific transitions.
- Thus, in a schooling example, Z could include family background variables while W_s could include college tuition or opportunity wages for unskilled labor.

- Collect the W_s into $W = (W_1, \ldots, W_{\bar{S}})$, and the U_s into $U = (U_1, \ldots, U_{\bar{S}})$.
- Larger values of $C_s(W_s)$ make it more likely that $D_s = 1$.
- The inequality restrictions on the $C_s(W_s)$ functions play a critical role in defining the model and producing its statistical implications.

- Analogous to the assumptions made for the binary outcome model, we assume

(OC-1)

$(U_s, V) \perp\!\!\!\perp (Z, W) | X$, $s = 1, \ldots, \bar{S}$. (**Conditional Independence of the Instruments**).

(OC-2)

*$\mu_D(Z)$ is a nondegenerate random variable conditional on X and W. (**Rank Condition**).*

(OC-3)

The distribution of V is continuous.

(OC-4)

$E(|Y_s|) < \infty$, $s = 1, \ldots, \bar{S}$. (**Finite Means**).

(OC-5)

$0 < \Pr(D_s = 1|X) < 1$ *for* $s = 1, \ldots, \bar{S}$ *for all* X. *(***In large samples, there are some persons in each treatment state***).*

(OC-6)

For $s = 1, \ldots, \bar{S} - 1$, the distribution of $C_s(W_s)$ conditional on X, Z and the other $C_j(W_j)$, $j = 1, \ldots, \bar{S}$ $j \neq s$, is nondegenerate and continuous.

- Assumption (OC-1) to (OC-5) play roles analogous to their counterparts in the two outcome model.
- (OC-6) is a new condition that is key to identification of the Δ^{MTE} defined below for each transition.
- It assumes that we can vary the choice sets of agents at different margins of schooling choice without affecting other margins of choice.
- A necessary condition for (OC-6) to hold is that at least one element of W_s is nondegenerate and continuous conditional on X, Z and $C_j(W_j)$ for $j \neq s$.

- Intuitively, one needs an instrument (or source of variability) for each transition.
- The continuity of the regressor allows us to differentiate with respect to $C_s(W_s)$, like we differentiated with respect to $P(Z)$ to estimate the MTE in the analysis of the two outcome model.

- The analysis of Angrist and Imbens (1995) discussed in the introduction to this section makes independence and monotonicity assumptions that generalize their earlier work.
- They do not consider estimation of transition-specific parameters as we do, or even transition-specific LATE.
- We present a different decomposition of the IV estimator where each component can be recovered from the data, and where the transition-specific MTEs answer well defined and economically interpretable policy evaluation questions.

- The probability of $D_s = 1$ given X, Z and W is generated by an ordered choice model:

$$\begin{aligned}\Pr\left(D_s = 1 \mid W, Z, X\right) &\equiv P_s(Z, W, X) \\ &= \Pr\left(C_{s-1}(W_{s-1}) < \mu_D(Z) - V \leq C_s(W_s) \mid X\right).\end{aligned}$$

- Analogous to the binary case, we can define $U_D = F_V(V|X = x)$ so $U_D \sim \text{Unif}[0, 1]$ under our assumption that the distribution of V is absolutely continuous with respect to Lebesgue measure.
- The probability integral transformation used extensively in the binary choice model is somewhat less useful for analyzing ordered choices, so we work with both U_D and V in this section of the paper.

- Monotonic transformations of V induce monotonic transformations of $\mu_D(Z) - C_s(W_s)$, but one is not free to form arbitrary monotonic transformations of $\mu_D(Z)$ and $C_s(W_s)$ separately.
- Using the probability integral transformation, the expression for choice s is $D_s = \mathbf{1}[F_V(\mu_D(Z) - C_{s-1}(W_{s-1})) > U_D \geq F_V(\mu_D(Z) - C_s(W_s))]$.
- Keeping the conditioning on X implicit, we define $P_s(Z, W) = F_V(\mu_D(Z) - C_{s-1}(W_{s-1})) - F_V(\mu_D(Z) - C_s(W_s))$.
- It is convenient to work with the probability that $S > s$, $\pi_s(Z, W_s) = F_V(\mu_D(Z) - C_s(W_s)) = \Pr\left(\sum_{j=s+1}^{\bar{S}} D_j = 1 \,\middle|\, Z, W_s \right)$, $\pi_{\bar{S}}(Z, W_{\bar{S}}) = 0$, $\pi_0(Z, W_0) = 1$ and $P_s(Z, W) = \pi_{s-1}(Z, W_{s-1}) - \pi_s(Z, W_s)$.

- The transition-specific Δ^{MTE} for the transition from s to $s+1$ is defined in terms of U_D.

$$\Delta^{\text{MTE}}_{s,s+1}(x, u_D) = E(Y_{s+1} - Y_s \mid X = x, U_D = u_D), \qquad s = 1, \dots, \bar{S} - 1.$$

- Alternatively, one can condition on V.
- Analogous to the analysis of the earlier sections of this paper, when we set $u_D = \pi_s(Z, W_s)$ we obtain the mean return to persons indifferent between s and $s+1$ at mean level of utility $\pi_s(Z, W_s)$.

- In this notation, keeping X implicit, the mean outcome Y, conditional on (Z, W), is the sum of the mean outcomes conditional on each state weighted by the probability of being in each state summed over all states:

$$\begin{aligned} E(Y|Z,W) &= \sum_{s=1}^{\bar{S}} E(Y_s \mid D_s = 1, Z, W) \Pr(D_s = 1 \mid Z, W) \quad (35) \\ &= \sum_{s=1}^{\bar{S}} \int_{\pi_s(Z,W_s)}^{\pi_{s-1}(Z,W_{s-1})} E(Y_s \mid U_D = u_D) du_D, \end{aligned}$$

where we use conditional independence assumption (OC-1) to obtain the final expression.

- Analogous to the result for the binary outcome model, we obtain the index sufficiency restriction $E(Y|Z,W) = E(Y \mid \pi(Z,W))$, where $\pi(Z,W) = [\pi_1(Z,W_1), \ldots, \pi_{\bar{S}-1}(Z,W_{\bar{S}-1})]$.
- The choice probabilities encode all of the influence of (Z,W) on outcomes.

- We can identify $\pi_s(z, w_s)$ for (z, w_s) in the support of the distribution of (Z, W_s) from the relationship $\pi_s(z, w_s) = \Pr(\sum_{j=s+1}^{\bar{S}} D_j = 1 \mid Z = z, W_s = w_s)$.
- Thus $E(Y \mid \pi(Z, W) = \pi)$ is identified for all π in the support of $\pi(Z, W)$.
- Assumptions (OC-1), (OC-3), and (OC-4) imply that $E(Y \mid \pi(Z, W) = \pi)$ is differentiable in π.

- So $\frac{\partial}{\partial \pi} E(Y \mid \pi(Z, W) = \pi)$ is well-defined.
- Thus analogous to the result obtained in the binary case

$$\begin{aligned} \frac{\partial E(Y \mid \pi(Z, W) = \pi)}{\partial \pi_s} &= \Delta^{\text{MTE}}_{s,s+1}(U_D = \pi_s) \qquad (36) \\ &= E(Y_{s+1} - Y_s \mid U_D = \pi_s). \end{aligned}$$

- Equation (36) is the basis for identification of the transition-specific MTE from data on (Y, Z, X).

- From index sufficiency, we can express (35) as

$$
\begin{aligned}
E\left(Y \mid \pi(Z,W)=\pi\right) &= \sum_{s=1}^{\bar{S}} E(Y_s \mid \pi_s \le U_D < \pi_{s-1})(\pi_{s-1}-\pi_s) \quad (37)\\
&= \sum_{s=1}^{\bar{S}-1} \left[\begin{array}{c} E(Y_{s+1} \mid \pi_{s+1} \le U_D < \pi_s) \\ -E(Y_s \mid \pi_s \le U_D < \pi_{s-1}) \end{array} \right] \pi_s \\
&\quad + E\left(Y_1 \mid \pi_1 \le U_D < 1\right) \\
&= \sum_{s=1}^{\bar{S}-1} \{ m_{s+1}(\pi_{s+1},\pi_s) - m_s(\pi_s,\pi_{s-1}) \} \pi_s \\
&\quad + E\left(Y_1 \mid \pi_1 \le U_D < 1\right)
\end{aligned}
$$

where $m_s(\pi_s,\pi_{s-1}) = E[Y_s \mid \pi_s \le U_D < \pi_{s-1}]$.

- In general this expression is a nonlinear function of (π_s, π_{s-1}).

- This model has a testable restriction of index sufficiency in the general case: $E(Y|\pi(Z,W)=\pi)$ is a nonlinear function that is additive in functions of (π_s, π_{s-1}) so there are no interactions between π_s and $\pi_{s'}$ if $|s-s'|>1$, i.e.,

$$\frac{\partial^2 E(Y \mid \pi(Z,W)=\pi)}{\partial \pi_s \partial \pi_{s'}} = 0 \qquad \text{if } |s-s'|>1.$$

- Observe that if $U_D \perp\!\!\!\perp U_s$ for $s = 1, \ldots, \bar{S}$,

$$
\begin{aligned}
E\left(Y \mid \pi(Z, W) = \pi\right) &= \sum_{s=1}^{\bar{S}} E(Y_s)(\pi_{s-1} - \pi_s) \\
&= \sum_{s=1}^{\bar{S}-1} \left[E(Y_{s+1}) - E(Y_s)\right] \pi_s + E(Y_1).
\end{aligned}
$$

  Defining $E(Y_{s+1}) - E(Y_s) = \Delta^{\text{ATE}}_{s,s+1}$,
  $E(Y \mid \pi(Z, W) = \pi) = \sum_{s=1}^{\bar{S}-1} \Delta^{\text{ATE}}_{s,s+1} \pi_s + E(Y_1)$.

- Thus, under full independence, we obtain linearity of the conditional mean of Y in the π_s's.

- This result generalizes the test for the presence of essential heterogeneity to the ordered case.
- We can ignore the complexity induced by the model of essential heterogeneity if $E(Y \mid \pi(Z, W) = \pi)$ is linear in the π's and can use conventional IV estimators to identify well-defined treatment effects.

What do Instruments Identify in the Ordered Choice Model?

- We now characterize what scalar instrument $J(Z, W)$ identifies.
- When Y is log earnings, it is common practice to regress Y on D where D is completed years of schooling and call the coefficient on D a rate of return.
- We seek an expression for the instrumental variables estimator of the effect of D on Y in the ordered choice model:

$$\frac{\mathrm{Cov}(J(Z, W), Y)}{\mathrm{Cov}(J(Z, W), D)}, \tag{38}$$

where $D = \sum_{s=1}^{\bar{S}} sD_s$ the number of years of schooling attainment.

- We keep the conditioning on X implicit.
- We now present the weights for IV.
- Define $K_s(v) =$
$E\left(\tilde{J}(Z, W) \mid \mu_D(Z) - c_s(W_s) > v\right) \Pr\left(\mu_D(Z) - C_s(W) > v\right)$,
where $\tilde{J}(Z, W) = J(Z, W) - E(J(Z, W))$.

- Thus,

$$\Delta_j^{\mathrm{IV}} = \frac{\mathrm{Cov}(J,Y)}{\mathrm{Cov}(J,D)} \tag{39}$$
$$= \sum_{s=1}^{\bar{S}-1} \int E(Y_{s+1} - Y_s \mid V = v)\omega(s,v)\, f_V(v)dv,$$

where

$$\omega(s,v) = \frac{K_s(v)}{\sum_{s=1}^{\bar{S}} s \int \left[K_{s-1}(v) - K_s(v)\right]\, f_V(v)dv}$$
$$= \frac{K_s\left(v\right)}{\sum_{s=1}^{\bar{S}-1} \int K_s\left(v\right) f_V\left(v\right) dv},$$

and clearly $\sum_{s=1}^{\bar{S}-1} \int \omega(s,v)\, f_V\left(v\right) dv = 1$, $\omega(0,v) = 0$, and $\omega(\bar{S},v) = 0$.

- We can rewrite this result in terms of the MTE, expressed in terms of u_D

$$\Delta^{\text{MTE}}_{s,s+1}(u_D) = E\left(Y_{s+1} - Y_s \mid U_D = u_D\right)$$

so that

$$\frac{\text{Cov}(J,Y)}{\text{Cov}(J,D)} = \sum_{s=1}^{\bar{S}-1} \int \Delta^{\text{MTE}}_{s,s+1}(u_D)\tilde{\omega}(s,u)\, du_D,$$

where

$$\begin{aligned} \tilde{\omega}(s,u_D) &= \frac{\tilde{K}_s(u_D)}{\sum_{s=1}^{\bar{S}} s \int_0^1 \left[\tilde{K}_{s-1}(u_D) - \tilde{K}_s(u_D)\right] du_D} && (40) \\ &= \frac{\tilde{K}_s\left(u_D\right)}{\sum_{s=1}^{\bar{S}-1} \int_0^1 \tilde{K}_s\left(u_D\right) du_D} \end{aligned}$$

and

$$\tilde{K}_s(u_D) = E\left(\tilde{J}(Z,W) \mid \pi_s(Z,W_s) > u_D\right) \Pr\left(\pi_s(Z,W_s) \geq u_D\right). \quad (41)$$

- The numerator of the weights for the Δ^{MTE} for a particular transition in the ordered choice model is exactly the numerator of the weights implied for the binary choice model, substituting $\pi_s(Z, W_s) = \Pr(D > s \mid Z, W_s)$ for $P(Z) = \Pr(D = 1 \mid Z)$.
- The numerator for the weights for IV in the binary choice model is driven by the connection between the instrument and $P(Z)$.
- The numerator for the weights for IV in the ordered choice model for a particular transition is driven by the connection between the instrument and $\pi_s(Z, W_s)$.

- The denominator of the weights is the covariance between the instrument and D for both the binary and ordered cases.
- However, in the binary case the covariance between the instrument and D is completely determined by the covariance between the instrument and $P(Z)$, while in the ordered choice case the covariance depends on the relationship between the instrument and the full vector $[\pi_1(Z, W_1), \ldots, \pi_{\bar{S}-1}(Z, W_{\bar{S}-1})]$.
- Comparing our decomposition of Δ^{IV} to decomposition (34), ours corresponds to weighting up marginal outcomes across well defined and adjacent boundary values experienced by agents having their instruments manipulated whereas the Angrist-Imbens decomposition corresponds to outcomes not experienced by some of the persons whose instruments are being manipulated.

- From equation (41), the IV estimator using $J(Z, W)$ as an instrument satisfies the following properties.
- (a) The numerator of the weights on $\Delta^{\text{MTE}}_{s,s+1}(u_D)$ is non-negative for all u_D if $E(J(Z, W_s) \mid \pi_s(Z, W_s) \geq \pi_s)$ is weakly monotonic in π_s.
- For example, if $\text{Cov}(\pi_s(Z, W_s), D) > 0$, setting $J(Z, W) = \pi_s(Z, W_s)$ will lead to nonnegative weights on $\Delta^{\text{MTE}}_{s,s+1}(u_D)$, though it may lead to negative weights on other transitions.

- A second property (b) is that the support of the weights on $\Delta^{\text{MTE}}_{s,s+1}$ using $\pi_s(Z, W_s)$ as the instrument is $(\pi_s^{\text{Min}}, \pi_s^{\text{Max}})$ where π_s^{Min} and π_s^{Max} are the minimum and maximum values in the support of $\pi_s(Z, W_s)$, respectively, and the support of the weights on $\Delta^{\text{MTE}}_{s,s+1}$ using any other instrument is a subset of $(\pi_s^{\text{Min}}, \pi_s^{\text{Max}})$.
- A third property (c) is that the weights on $\Delta^{\text{MTE}}_{s,s+1}$ implied by using $J(Z, W)$ as an instrument are the same as the weights on $\Delta^{\text{MTE}}_{s,s+1}$ implied by using $E(J(Z, W) \mid \pi_s(Z, W))$ as the instrument.

- Suppose that the distributions of W_s, $s = 1, \ldots, \bar{S}$, are degenerate so that the C_s are constants satisfying $C_1 < \cdots < C_{\bar{S}-1}$.
- This is the classical ordered choice model.
- In this case, $\pi_s(Z, W_s) = F_V(\mu_D(Z) - C_s)$ for any $s = 1, \ldots, \bar{S}$.
- For this special case, using J as an instrument will lead to nonnegative weights on all transitions if $J(Z, W_s)$ is a monotonic function of $\mu_D(Z)$.
- For example, note that $\mu_D(Z) - C_s > v$ can be written as $\mu_D(Z) > C_s + F_V^{-1}(u_D)$.

- Using $\mu_D(Z)$ as the instrument leads to weights on $\Delta^{\text{MTE}}_{s,s+1}(u_D)$ of the form specified above with

$$\tilde{K}_s(u_D) = \Big[E(\mu_D(Z) \mid \mu_D(Z) > F_V^{-1}(u_D) + C_s) - E(\mu_D(Z))\Big] \Pr(\mu_D(Z) > F_V^{-1}(u_D) + C_s).$$

- Clearly, these weights will be nonnegative for all points of evaluation and will be strictly positive for any evaluation point u_D such that $1 > \Pr(\mu_D(Z) > F_V^{-1}(u_D) + C_s) > 0$.
- We now present some examples of the weights for IV.

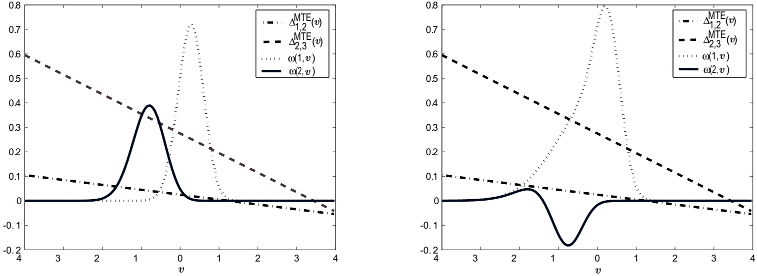
Examples of Weights for IV

- Figures 1 and 2 plot the transition-specific MTEs and the IV weights for the models and distributions of the data at the base of each of the figures.
- We work with a normal V and U_s, so we get linear in V MTEs from standard normal regression theory.
- The IV estimates using Z and W_1 as instruments are reported transition by transition, along with the overall IV representation (39) into its transition-specific components.
- The IV weights are defined by equations (40) and (41). The bottom table presents the transition-specific treatment parameters.

Figure 1:

The Generalized Ordered Choice Roy Model under Normality: Case I

A. Z as Instrument B. W_1 as Instrument



Outcomes

$$Y_1 = \alpha + \beta_1 + U_1$$
$$Y_2 = \alpha + \beta_2 + U_2$$
$$Y_3 = \alpha + \beta_3 + U_3$$

Choice Model

$$D_s = \mathbf{1}[W_{s-1} < \gamma Z - V \leqslant W_s]$$
$$s = 1, 2, 3$$

Parameterization

$$(U_1, U_2, U_3, V) \sim N\left(\mathbf{0}, \mathbf{\Sigma}_{UV}\right),\ (Z, W_1, W_2) \sim N\left(\mu_{ZW}, \mathbf{\Sigma}_{ZW}\right) \text{ and } W_0 = -\infty; W_3 = \infty.$$

$$\mathbf{\Sigma}_{UV} = \begin{bmatrix} 1 & 0.16 & 0.2 & -0.3 \\ 0.16 & 0.64 & 0.16 & -0.32 \\ 0.2 & 0.16 & 1 & -0.4 \\ -0.3 & -0.32 & -0.4 & 1 \end{bmatrix},\ \mu_{ZW} = (-0.6, -1.08, 0.08) \text{ and } \mathbf{\Sigma}_{ZW} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & -0.09 \\ 0 & -0.09 & 0.25 \end{bmatrix}$$

$$\mathrm{Cov}(U_2 - U_1, V) = -0.02 \quad \mathrm{Cov}(U_3 - U_2, V) = -0.08$$

$$\beta_1 = 0;\ \beta_2 = 0.025;\ \beta_3 = 0.3;\ \gamma = 1$$

IV Estimates and Their Components*

Parameter	Value
Δ^{IV_Z}	0.1489
$\Delta_{12}^{\mathrm{IV}_Z}$	*0.0117*
$\Delta_{23}^{\mathrm{IV}_Z}$	*0.1372*
$\Delta^{\mathrm{IV}_{W_1}}$	0.0017
$\Delta_{12}^{\mathrm{IV}_{W_1}}$	*0.0325*
$\Delta_{23}^{\mathrm{IV}_{W_1}}$	*−0.0308*

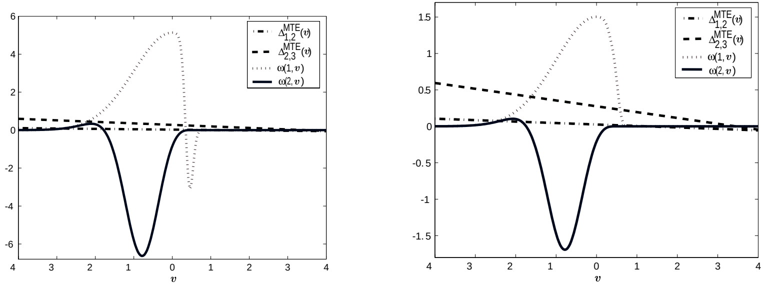
Treatment Parameters and Their Values

Parameter	Value
$\mathrm{ATE}_{12} = E\left(Y_2 - Y_1\right)$	0.025
$\mathrm{ATE}_{23} = E\left(Y_3 - Y_2\right)$	0.275
$\mathrm{TT}_{12} = E\left(Y_2 - Y_1 \mid D_2 = 1\right)$	0.0271
$\mathrm{TT}_{23} = E\left(Y_3 - Y_2 \mid D_3 = 1\right)$	0.1871
$\mathrm{TUT}_{12} = E\left(Y_2 - Y_1 \mid D_1 = 1\right)$	0.0047
$\mathrm{TUT}_{23} = E\left(Y_3 - Y_2 \mid D_2 = 1\right)$	0.2854

Figure 2:

The Generalized Ordered Choice Roy Model under Normality: Case II

A. Z as Instrument B. W_1 as Instrument



Outcomes

$$Y_1 = \alpha + \beta_1 + U_1$$
$$Y_2 = \alpha + \beta_2 + U_2$$
$$Y_3 = \alpha + \beta_3 + U_3$$

Choice Model

$$D_s = \mathbf{1}[W_{s-1} < \gamma Z - V \leqslant W_s]$$
$$s = 1, 2, 3$$

Parameterization

$$(U_1, U_2, U_3, V) \sim N(\mathbf{0}, \mathbf{\Sigma}_{UV}), \quad (Z, W_1, W_2) \sim N(\mu_{ZW}, \mathbf{\Sigma}_{ZW}) \quad \text{and} \quad W_0 = -\infty; W_3 = \infty.$$

$$\mathbf{\Sigma}_{UV} = \begin{bmatrix} 1 & 0.16 & 0.2 & -0.3 \\ 0.16 & 0.64 & 0.16 & -0.32 \\ 0.2 & 0.16 & 1 & -0.4 \\ -0.3 & -0.32 & -0.4 & 1 \end{bmatrix}, \ \mu_{ZW} = (-0.6, -1.08, 0.08) \text{ and } \mathbf{\Sigma}_{ZW} = \begin{bmatrix} 0.1 & 0.092 & -0.036 \\ 0.092 & 0.1 & -0.09 \\ -0.036 & -0.09 & 0.25 \end{bmatrix}$$

$$\text{Cov}(U_2 - U_1, V) = -0.02 \quad \text{Cov}(U_3 - U_2, V) = -0.08$$

$$\beta_1 = 0;\ \beta_2 = 0.025;\ \beta_3 = 0.3;\ \gamma = 1$$

IV Estimates and Their Components†

Parameter	Value
Δ^{IV_Z}	-1.8091
$\Delta_{12}^{\text{IV}_Z}$	*0.2866*
$\Delta_{23}^{\text{IV}_Z}$	*-2.0957*
$\Delta^{\text{IV}_{W_1}}$	-0.4284
$\Delta_{12}^{\text{IV}_{W_1}}$	*0.0909*
$\Delta_{23}^{\text{IV}_{W_1}}$	*-0.5193*

Treatment Parameters and Their Values

Parameter	Value
$\text{ATE}_{12} = E(Y_2 - Y_1)$	0.025
$\text{ATE}_{23} = E(Y_3 - Y_2)$	0.275
$\text{TT}_{12} = E(Y_2 - Y_1 \mid D_2 = 1)$	0.0283
$\text{TT}_{23} = E(Y_3 - Y_2 \mid D_3 = 1)$	0.1754
$\text{TUT}_{12} = E(Y_2 - Y_1 \mid D_1 = 1)$	0.0025
$\text{TUT}_{23} = E(Y_3 - Y_2 \mid D_2 = 1)$	0.2898

- In Figure 1, the IV weights based on Z and W_1 are very different.
- So, correspondingly, are the IV estimates produced from each instrument, which are far off the mark of the standard treatment parameters shown at the bottom of the table.
- Observe that the IV weight for W_1 in the second transition is negative for an interval of values.
- This accounts for the dramatically lower IV estimate based on W_1 as the instrument.
- Figure 2 shows a different configuration of (Z, W_1, W_2).

- This produces negative weights for Z for both transitions and a negative weight for W_1 in the second transition.
- For both instruments, IV is negative even though both MTEs are positive throughout most of their range.
- IV provides a misleading summary of the underlying marginal treatment effects.
- In digesting Figures 1 and 2, it is important to recall that all are based on the same structural model.
- All have the same MTE and average treatment effects.
- But the IV estimates are very different solely as a consequence of the differences in the distributions of instruments across examples.

- These simulations show a rich variety of shapes and signs for the weights.
- They illustrate a main point of this paper—that standard IV methods are not guaranteed to weight marginal treatment effects positively or to produce estimates close to any of the standard treatment effects.
- Estimators based on LIV and its extension to the ordered model (36) identify Δ^{MTE} for each transition and answer policy relevant questions.
- We now turn to development of a more general unordered model.

Extension to Multiple Treatments that are Unordered

- In this section, we develop a framework for multiple treatments with a choice equation that is based on a nonparametric version of the classical multinomial choice model.
- Within this framework, treatment effects can be defined as the difference in the counterfactual outcomes that would have been observed if the agent faced different choice sets, i.e., the effect of the individual being forced to choose from one choice set instead of another.

- We analyze the return to the agent of choosing between option j and the next best option.
- The analysis of this case is very similar because it converts a multiple choice problem to a binary choice problem.
- Exclusion restrictions allow analysts to identify generalizations of the LATE parameter and MTE parameters corresponding to the effect of one choice versus the "next-best" alternative.
- This identification analysis does not require large support assumptions.

- Consider the following model with multiple outcome states.
- Let $\mathcal{J}$ denote the agent's choice set, where $\mathcal{J}$ contains a finite number of elements.
- The reward (psychic and monetary) of choosing $j \in \mathcal{J}$ is

$$R_j(Z_j) = \vartheta_j(Z_j) - V_j, \tag{42}$$

where Z_j are the agent's observed characteristics that affect the utility from choosing choice j, and V_j is the unobserved shock to the agent's utility from choice j.

- Let Z denote the random vector containing all unique elements of $\{Z_j\}_{j\in\mathcal{J}}$, i.e., $Z =$ union of $\{Z_j\}_{j\in\mathcal{J}}$.
- We write $R_j(Z)$ for $R_j(Z_j)$, leaving implicit that $R_j(\cdot)$ only depends on those elements of Z that are contained in Z_j.
- Let $D_{\mathcal{J},j}$ be an indicator variable for whether the agent would choose option j if confronted with choice set $\mathcal{J}$:

$$D_{\mathcal{J},j} = \begin{cases} 1 & \text{if } R_j \geq R_k \quad \forall k \in \mathcal{J} \\ 0 & \text{otherwise.} \end{cases}$$

- Let $I_{\mathcal{J}}$ denote the choice that would be made by the agent if confronted with choice set $\mathcal{J}$: $I_{\mathcal{J}} = j \iff D_{\mathcal{J},j} = 1$.
- Let $Y_{\mathcal{J}}$ be the outcome variable that would be observed if the agent faced choice set $\mathcal{J}$.
- It is

$$Y_{\mathcal{J}} = \sum_{j \in \mathcal{J}} D_{\mathcal{J},j} Y_j, \tag{43}$$

where Y_j is the potential outcome, observed only if option j is chosen.

- We assume that Y_j is determined by $Y_j = \mu_j(X_j, U_j)$, where X_j is a vector of the agent's observed characteristics and U_j is an unobserved random vector.
- Let X denote the random vector containing all unique elements of $\{X_j\}_{j \in \mathcal{J}}$, i.e., X is the union of $\{X_j\}_{j \in \mathcal{J}}$.
- We assume that $(Z, X, I_{\mathcal{J}}, Y_{\mathcal{J}})$ is observed.

- Define $R_{\mathcal{J}}$ as the maximum obtainable value given choice set $\mathcal{J}$:

$$R_{\mathcal{J}} = \max_{j \in \mathcal{J}} \{R_j\} = \sum_{j \in \mathcal{J}} D_{\mathcal{J},j} R_j.$$

- We obtain the traditional representation of the decision process that if choice j is optimal, choice j is better than the "next best" option:

$$I_{\mathcal{J}} = j \iff R_j \geq R_{\mathcal{J} \setminus j},$$

where $\mathcal{J} \setminus j$ means $\mathcal{J}$ removing the j^{th} element from the set.

- More generally, a choice with $\mathcal{K}$ optimal is equivalent to the highest value obtainable from choices in $\mathcal{K}$ being higher than the highest value that can be obtained from choices outside that set,

$$I_{\mathcal{J}} \in \mathcal{K} \Longleftrightarrow R_{\mathcal{K}} \geq R_{\mathcal{J} \setminus \mathcal{K}}.$$

- As we will show, this well-known representation used by Lee (1983), Dahl (2002) and others, is key for understanding how nonparametric instrumental variables estimates the effect of a given choice versus the "next best" alternative.

- Analogous to our definition of $R_{\mathcal{J}}$, we define $R_{\mathcal{J}}(z)$ to be the maximum attainable value given choice set $\mathcal{J}$ when instruments are fixed at $Z = z$,

$$R_{\mathcal{J}}(z) = \max_{j \in \mathcal{J}} \{R_j(z)\}.$$

- Thus, for example, a choice from $\mathcal{K}$ is optimal when instruments are fixed at $Z = z$ if $R_{\mathcal{K}}(z) \geq R_{\mathcal{J} \setminus \mathcal{K}}(z)$.

- We make the following assumptions, which generalize assumptions for the multiple treatment case and are presented in a parallel fashion ((B-2) is stated below):

(B-1)

$\{(V_j, U_j)\}_{j \in \mathcal{J}}$ *is independent of* Z *conditional on* X.

(B-3)

The distribution of $(\{V_j\}_{j\in\mathcal{J}})$ *is absolutely continuous with respect to Lebesgue measure on* $\prod_{j\in\mathcal{J}}\mathbb{R}$.

(B-4)

$E|Y_j| < \infty$ *for all* $j \in \mathcal{J}$.

(B-5)

$Pr(I_{\mathcal{J}} = j|X) > 0$ *for all* $j \in \mathcal{J}$.

- Assumptions (B-1) and (B-3) imply that $R_j \neq R_k$ w.p.1 for $j \neq k$, so that argmax$\{R_j\}$ is unique w.p.1.
- Assumption (B-4) is required for the mean treatment parameters to be well defined.
- Assumption (B-5) requires that at least some individuals participate in each program for all X.

- Definitions of the treatment parameters only require assumptions (B-1) and (B-3) to (B-5). However, we use exclusion restrictions to secure identification.
- Let $Z^{[j]}$ denote the j th component of Z.
- Let $Z^{[-j]}$ denote all elements of Z except for the jth component.
- We will work with two alternative assumptions for the exclusion restriction.

- Consider

(B-2a) For each $j \in \mathcal{J}$, their exists at least one element of Z, say $Z^{[j]}$, such that $Z^{[j]}$ is not an element of Z_k, $k \neq j$, and such that the distribution of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-j]})$ is nondegenerate,

or

(B-2b) For each $j \in \mathcal{J}$, their exists at least one element of Z, say $Z^{[j]}$, such that $Z^{[j]}$ is not an element of Z_k, $k \neq j$, and such that the distribution of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-j]})$ is absolutely continuous with respect to Lebesgue measure.

- Assumption (B-2a) requires that the analyst be able to independently vary the index for the given value function.
- It imposes an exclusion restriction, that for any $j \in \mathcal{J}$, Z contains an element such that (i) it is contained in Z_j; (ii) it is not contained in any Z_k for $k \neq j$, and (iii) $\vartheta_j(\cdot)$ is a nontrivial function of that element conditional on all other regressors.
- Assumption (B-2b) strengthens (B-2a) by adding a smoothness assumption.
- A necessary condition for (B-2b) is that the excluded variable have a density with respect to Lebesgue measure conditional on all other regressors and for $\vartheta_j(\cdot)$ to be a continuous and nontrivial function of the excluded variable.

- Assumption (B-2a) is used to identify a generalization of the LATE parameter.
- Assumption (B-2b) will be used to identify a generalization of the MTE parameter.
- Below, we will strengthen (B-2b) to a large support assumption to identify ATE though the large support assumption will not be required for most of our analysis.
- Assumptions (B-2a) and (B-2b) are analogous to (OC-2) and (OC-6) in an ordered choice setting.

Definition of Treatment

- Treatment effects are defined as the difference in the counterfactual outcomes that would have been observed if the agent faced different choice sets.
- For any two choice sets, $\mathcal{K}, \mathcal{L} \subset \mathcal{J}$, define $\Delta_{\mathcal{K},\mathcal{L}} = Y_{\mathcal{K}} - Y_{\mathcal{L}}$, the effect of the individual being forced to choose from choice set $\mathcal{K}$ versus choice set $\mathcal{L}$.

- The conventional treatment effect is defined as the difference in potential outcomes between two specified states,

$$\Delta_{k,\ell} = Y_k - Y_\ell,$$

  which is nested within this framework by taking $\mathcal{K} = \{k\}$, $\mathcal{L} = \{\ell\}$.

- It is the effect for the individual of having no choice except to choose state k versus having no choice except to choose state ℓ.

- $\Delta_{\mathcal{K},\mathcal{L}}$ will be zero for agents who make the same choice when confronted with choice set $\mathcal{K}$ and choice set $\mathcal{L}$.
- Thus, $I_{\mathcal{K}} = I_{\mathcal{L}}$ implies $\Delta_{\mathcal{K},\mathcal{L}} = 0$, and we have

$$\begin{aligned} \Delta_{\mathcal{K},\mathcal{L}} &= \mathbf{1}(I_{\mathcal{L}} \neq I_{\mathcal{K}})\Delta_{\mathcal{K}\setminus\mathcal{L},\mathcal{L}} \qquad (44) \\ &= \mathbf{1}(I_{\mathcal{L}} \neq I_{\mathcal{K}}) \left(\sum_{j \in \mathcal{K}\setminus\mathcal{L}} D_{\mathcal{K},j}\Delta_{j,\mathcal{L}} \right). \end{aligned}$$

- Two cases will be of particular importance for our analysis.
- First, consider choice set $\mathcal{K} = \{k\}$ versus choice set $\mathcal{L} = \mathcal{J} \setminus \{k\}$.
- In this case, $\Delta_{k,\mathcal{J}\setminus k}$ is the difference between the agent's potential outcome in state k versus the outcome that would have been observed if he or she had not been allowed to choose state k.
- If $I_{\mathcal{J}} = k$, then $\Delta_{k,\mathcal{J}\setminus k}$ is the difference between the outcome in the agent's preferred state and the outcome in the agent's "next-best" state.

- Second, consider the set $\mathcal{K} = \mathcal{J}$ versus choice set $\mathcal{L} = \mathcal{J} \setminus \{k\}$. In this case, $\Delta_{\mathcal{J},\mathcal{J}\setminus k}$ is the difference between the agent's observed outcome and what his or her outcome would have been if state k had not been available.
- Note that $\Delta_{\mathcal{J},\mathcal{J}\setminus k} = D_{\mathcal{J},k}\Delta_{k,\mathcal{J}\setminus k}$.
- Thus, there is a trivial connection between the two parameters, $\Delta_{\mathcal{J},\mathcal{J}\setminus k}$ and $\Delta_{k,\mathcal{J}\setminus k}$.
- This paper focuses on $\Delta_{k,\mathcal{J}\setminus k}$, the effect of being forced to choose option k versus being denied option k.
- However, one can exploit equation (44) to use the results for $\Delta_{k,\mathcal{J}\setminus k}$ to obtain results for $\Delta_{\mathcal{J},\mathcal{J}\setminus k}$.

Treatment Parameters

- The conventional definition of the average treatment effect (ATE) is $\Delta^{\text{ATE}}_{k,\ell}(x,z) = E(\Delta_{k,\ell}|X = x, Z = z)$, which immediately generalizes to the class of parameters just discussed: $\Delta^{\text{ATE}}_{\mathcal{K},\mathcal{L}}(x,z) = E(\Delta_{\mathcal{K},\mathcal{L}}|X = x, Z = z)$.
- The conventional definition of the treatment on the treated (TT) parameter is $\Delta^{\text{TT}}_{k,\ell}(x,z) = E(\Delta_{k,\ell}|X = x, Z = z, I_{\mathcal{J}} = k)$, which generalizes to $\Delta^{\text{TT}}_{\mathcal{K},\mathcal{L}}(x,z) = E(\Delta_{\mathcal{K},\mathcal{L}}|X = x, Z = z, I_{\mathcal{J}} \in \mathcal{K})$.

- We generalize the MTE parameter to be the average effect conditional on being indifferent between the best option among choice set $\mathcal{K}$ versus the best option among choice set $\mathcal{L}$ at some fixed value of the instruments, $Z = z$:

$$\Delta^{\text{MTE}}_{\mathcal{K},\mathcal{L}}(x, z) = E\left(\Delta_{\mathcal{K},\mathcal{L}} \mid X = x, Z = z, R_{\mathcal{K}}(z) = R_{\mathcal{L}}(z)\right). \tag{45}$$

- We generalize the LATE parameter to be the average effect for someone for whom the optimal choice in choice set $\mathcal{K}$ is preferred to the optimal choice in choice set $\mathcal{L}$ at $Z = \tilde{z}$, but who prefers the optimal choice in choice set $\mathcal{L}$ to the optimal choice in choice set $\mathcal{K}$ at $Z = z$:

$$\Delta^{\text{LATE}}_{\mathcal{K},\mathcal{L}}(x, z, \tilde{z}) = E\left(\Delta_{\mathcal{K},\mathcal{L}} \;\middle|\; \begin{array}{c} X = x, Z \in \{z, \tilde{z}\}, R_{\mathcal{K}}(\tilde{z}) \geq R_{\mathcal{L}}(\tilde{z}), \\ R_{\mathcal{L}}(z) \geq R_{\mathcal{K}}(z) \end{array}\right). \tag{46}$$

- An important special case of this parameter arises when $z = \tilde{z}$ except for elements that enter the index functions only for choices in $\mathcal{K}$ and not for any choice in $\mathcal{L}$.
- In that special case, expression (46) simplifies to

$$\Delta^{\text{LATE}}_{\mathcal{K},\mathcal{L}}(x, z, \tilde{z}) = E\left(\Delta_{\mathcal{K},\mathcal{L}} \;\middle|\; \begin{array}{c} X = x, Z \in \{z, \tilde{z}\}, \\ R_{\mathcal{K}}(\tilde{z}) \geq R_{\mathcal{L}}(z) \geq R_{\mathcal{K}}(z) \end{array}\right)$$

since $R_{\mathcal{L}}(z) = R_{\mathcal{L}}(\tilde{z})$ in this special case.

- We have defined each of these parameters as conditional not only on X but also on the "instruments" Z.
- In general, the parameters will depend on the Z evaluation point.
- For example, $\Delta^{\text{ATE}}_{\mathcal{K},\mathcal{L}}(x,z)$ will in general depend on the z evaluation point.
- To see this, note that $Y_{\mathcal{K}} = \sum\limits_{k \in \mathcal{K}} D_{\mathcal{K},k} Y_k$, and $Y_{\mathcal{L}} = \sum\limits_{\ell \in \mathcal{L}} D_{\mathcal{L},\ell} Y_{\ell}$.

- By independence assumption (B-1), we have that $Z \perp\!\!\!\perp \{Y_j\}_{j \in \mathcal{J}} \mid X$, but $D_{\mathcal{K},k}$ and $D_{\mathcal{L},\ell}$ will be dependent on Z conditional on X and thus $Y_{\mathcal{K}} - Y_{\mathcal{L}}$ will in general be dependent on Z conditional on X.
- In other words, even though Z is conditionally independent of each individual potential outcome, it is correlated with which choice is optimal within the sets $\mathcal{K}$ and $\mathcal{L}$ and thus is related to $Y_{\mathcal{K}} - Y_{\mathcal{L}}$.

Identification: Effect of Option j Versus Next Best Alternative

- We now establish identification of treatment parameters corresponding to averages of $\Delta_{j,\mathcal{J}\setminus j}$, the effect of choosing option j versus the preferred option in $\mathcal{J}$ if j were not available.
- Recall that $Z^{[j]}$ is the vector of elements of Z_j that do not enter any other choice index, and that $Z^{[-j]}$ is a vector of all elements of Z not in $Z^{[j]}$.
- The $Z^{[j]}$ thus act as shifters attracting people into or out of j, but not affecting the valuations in the arguments of the other choice functions.

- We can develop a parallel analysis to the binary case developed earlier in this paper if we condition on $Z^{[-j]}$.
- We obtain monotonicity or uniformity in this model if the movements among states induced by $Z^{[j]}$ are the same for all persons conditional on $Z^{[-j]} = z^{[-j]}$ and $X = x$.
- For example, *ceteris paribus* if $Z^{[j]} = z^{[j]}$ increases, $R_j\left(Z_j\right)$ increases but the $R_k\left(Z_k\right)$ are not affected, so the flow is toward state j.

- Let $D_{\mathcal{J},j}$ be an indicator variable denoting whether option j is selected.

-

$$
\begin{aligned}
D_{\mathcal{J},j} &= \mathbf{1}\left(R_j(Z_j) \geq \max_{\ell \neq j}\{R_\ell(Z_\ell)\}\right) \\
&= \mathbf{1}\left(\vartheta_j(Z_j) \geq V_j + \max_{\ell \neq j}\{R_\ell(Z_\ell)\}\right) \\
&= \mathbf{1}\left(\vartheta_j(Z_j) \geq \tilde{V}_j\right),
\end{aligned} \tag{47}
$$

where $\tilde{V}_j = V_j + \max_{\ell \neq j}\{R_\ell(Z_\ell)\}$.

- Thus we obtain $D_{\mathcal{J},j} = \mathbf{1}\left(P_j(Z_j) \geq U_{D_j}\right)$, where $U_{D_j} = F_{\tilde{V}_j}(V_j + \max_{\ell \neq j}\{R_\ell(Z_\ell)\} \mid Z^{[-j]} = z^{[-j]})$, where $F_{\tilde{V}_j}$ is the cdf of $\tilde{V}_j$ given $Z^{[-j]} = z^{[-j]}$.

- In a format parallel to the binary model, we write

$$Y = D_{\mathcal{J},j} Y_j + (1 - D_{\mathcal{J},j}) Y_{\mathcal{J} \setminus j}, \tag{48}$$

where $Y_{\mathcal{J} \setminus j}$ is the outcome that would be observed if option j were not available.

- This case is just a version of the binary case developed in previous sections of the paper.

- We can define MTE as

$$E\left(Y_j - Y_{\mathcal{J} \setminus j} \mid X = x, Z = z, \vartheta_j(z_j) - V_j = R_{\mathcal{J} \setminus j}(z)\right).$$

Recall that we have to condition on $Z = z$ because the choice sets are defined over the max of elements in $\mathcal{J} \setminus j$ (see equation (47)).

- We now show that our identification strategies presented in the preceding part of this paper extend naturally to the identification of treatment parameters for $\Delta_{j,\mathcal{J}\setminus j}$.
- In particular, it is possible to recover LATE and MTE parameters for $\Delta_{j,\mathcal{J}\setminus j}$ by use of discrete change IV methods and local instrumental variable methods, respectively.
- Averages of the effect of option j versus the next best alternative are the easiest effects to study using instrumental variable methods and are natural generalizations of our two outcome analysis.

- Consider identification of treatment parameters corresponding to averages of $\Delta_{j,\mathcal{J}\setminus j}$ using either a discrete change, Wald form for the instrumental variables estimand or using the local instrumental variables (LIV) estimand.
- The discrete change, instrumental variables estimand will allow us to recover a version of the local average treatment effect (LATE) parameter.
- Let $Z^{[-j]}$ denote the excluded variable for option j with properties assumed in (B-2a). We let $z = \left[z^{[-j]}, z^{[j]}\right]$ and $\tilde{z} = \left[\tilde{z}^{[-j]}, \tilde{z}^{[j]}\right]$ be two values of Z where we only manipulate $Z^{[j]}$.

- Define

$$\Delta_j^{\text{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]}) = \frac{E(Y|X=x, Z=\tilde{z}) - E(Y|X=x, Z=z)}{\Pr(D_{\mathcal{J},j}=1|X=x, Z=\tilde{z}) - \Pr(D_{\mathcal{J},j}=1|X=x, Z=z)},$$

where for notational convenience we assume that $Z^{[j]}$ is the last component of Z.

- Without loss of generality, we assume that $\vartheta_j(\tilde{z}) > \vartheta_j(z)$.
- The local instrumental variables estimator (LIV) estimand introduced in Heckman (1997), and developed further in Heckman and Vytlacil (1999, 2001) allows us to recover a version of the Marginal Treatment Effect (MTE) parameter.

- Impose (B-2b), and let $Z^{[j]}$ denote the excluded variable for option j with properties assumed in (B-2b). Our results are invariant to which particular variable satisfying (B-2b) is used if there are more than one variable with the property assumed in (B-2b). Define

$$\Delta_j^{\text{LIV}}(x,z) \equiv \frac{\frac{\partial}{\partial z^{[j]}} E(Y \mid X = x, Z = z)}{\frac{\partial}{\partial z^{[j]}} \Pr(D_{\mathcal{J},j} = 1 \mid X = x, Z = z)}. \tag{49}$$

- $\Delta_j^{\text{LIV}}(x,z)$ is thus the limit form of $\Delta_j^{\text{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ as $\tilde{z}^{[j]}$ approaches $z^{[j]}$.
- Given our previous assumptions, one can easily show that this limit exists w.p.1.
- We prove the following identification theorem.

Theorem

1. *Assume (B-1), (B-3) to (B-5) and (B-2a). Then* $\Delta_j^{Wald}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]}) = \Delta_{j,\mathcal{J}\setminus j}^{LATE}(x, z, \tilde{z})$ *where* $\tilde{z} = (z^{[-j]}, \tilde{z}^{[j]})$.
2. *Assume (B-1), (B-3) to (B-5) and (B-2b). Then* $\Delta_j^{LIV}(x, z) = \Delta_{j,\mathcal{J}\setminus j}^{MTE}(x, z)$.

- The intuition underlying the proof is simple.
- Under (B-1), (B-3) to (B-5) and (B-2a) we can convert the problem of comparing the outcome under j with the outcome under the next best option.
- This is an IV version of the selection modelling analysis of Dahl (2002). $\Delta^{\text{LATE}}_{j,\mathcal{J}\setminus j}(x, z, \tilde{z})$ is the average effect of switching to state j from state $I_{\mathcal{J}\setminus j}$ for individuals who would choose $I_{\mathcal{J}\setminus j}$ at $Z = z$ but would choose j at $Z = \tilde{z}$.

- $\Delta^{\text{MTE}}_{j,\mathcal{J}\setminus j}(x,z)$ is the average effect of switching to state j from state $I_{\mathcal{J}\setminus j}$ (the best option besides state j) for individuals who are indifferent between state j and $I_{\mathcal{J}\setminus j}$ at the given values of the selection indices (at $Z=z$, i.e., at $\{\vartheta_k(Z_k)=\vartheta_k(z_k)\}_{k\in\mathcal{J}}$).
- The mean outcome in state j versus state $I_{\mathcal{J}\setminus j}$ (the next best option) is a weighted average over $k\in\mathcal{J}\setminus j$ of the effect of state j versus state k, conditional on k being the next best option, weighted by the probability that k is the next best option.

- For example, for the LATE parameter,

$$\begin{aligned}
\Delta^{\text{LATE}}_{j,\mathcal{J}\setminus j}(x,z,\tilde{z}) &= E\left(\Delta_{j,\mathcal{J}\setminus j} \;\middle|\; \begin{array}{l} X=x, Z\in\{z,\tilde{z}\}, \\ R_j(\tilde{z}) \geq R_{\mathcal{J}\setminus j}(z) \geq R_j(z) \end{array}\right) \\
&= \sum_{k\in\mathcal{J}\setminus j}\left[\begin{array}{l} \Pr\left(I_{\mathcal{J}\setminus j}=k \;\middle|\; \begin{array}{l} Z\in\{z,\tilde{z}\}, \\ R_j(\tilde{z}) \geq R_{\mathcal{J}\setminus j}(z) \geq R_j(z) \end{array}\right) \\ \times E\left(\Delta_{j,k} \;\middle|\; \begin{array}{l} X=x, Z\in\{z,\tilde{z}\}, \\ R_j(\tilde{z}) \geq R_{\mathcal{J}\setminus j}(z) \geq R_j(z), \\ I_{\mathcal{J}\setminus j}=k \end{array}\right) \end{array}\right],
\end{aligned}$$

where we use the fact that $R_{\mathcal{J}\setminus j}(z) = R_{\mathcal{J}\setminus j}(\tilde{z})$ since $z=\tilde{z}$ except for one component that only enters the index for the jth option.

- How heavily each option is weighted in this average depends on

$$\Pr\left(I_{\mathcal{J}\setminus j} = k \mid Z \in \{z, \tilde{z}\}, R_j(\tilde{z}_j) \geq R_k(z_k) \geq R_j(z_j)\right),$$

which in turn depends on $\{\vartheta_k(z_k)\}_{k \in \mathcal{J}\setminus j}$.

- The higher $\vartheta_k(z_k)$, holding the other indices constant, the larger the weight given to state k as the base state.
- The LIV and Wald estimands depend on the z evaluation point.

- Alternatively, one can define averaged versions of the LIV and Wald estimands that will recover averaged versions of the MTE and LATE parameters,

$$
\begin{aligned}
& \int \Delta_j^{\text{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]}) dF_{Z^{[-j]}}(z^{[-j]}) \\
= & \int \Delta_{j,\mathcal{J}\setminus j}^{\text{LATE}}(x, z, \tilde{z}) dF_{Z^{[-j]}}(z^{[-j]}) \\
= & E\left(\Delta_{j,\mathcal{J}\setminus j} \middle| \begin{array}{c} X = x, \\ R_j(Z^{[-j]}, \tilde{z}^{[j]}) \\ \geq R_{\mathcal{J}\setminus j}(Z^{[-j]}) \geq R_j(Z^{[-j]}, z^{[j]}) \end{array}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
\int \Delta_j^{\text{LIV}}(x, z) dF_Z(z) &= \int \Delta_{j,\mathcal{J}\setminus j}^{\text{MTE}}(x, z) dF_Z(z) \\
&= E\left(\Delta_{j,\mathcal{J}\setminus j} | X = x, R_j(Z) = R_{\mathcal{J}\setminus j}(Z)\right).
\end{aligned}
$$

- Thus far, we have only considered identification of LATE and MTE, and not of the more standard treatment parameters ATE and TT.
- However, following Heckman and Vytlacil (1999), LATE can approximate ATE or TT arbitrarily well given the appropriate support conditions.
- Theorem 3 shows that we can use Wald estimands to identify LATE for $\Delta_{j,\mathcal{J}\setminus j}$, and we can thus adapt Heckman and Vytlacil (1999) to identify ATE or TT for $\Delta_{j,\mathcal{J}\setminus j}$.
- With suitable modification of the weights, their analysis goes through as before.

- Suppose that $Z^{[j]}$ satisfies the properties assumed in (B-2a), and suppose that: (i) the support of the distribution of $Z^{[j]}$ conditional on all other elements of Z is the full real line; (ii) $\vartheta_j(z_j) \rightarrow \infty$ as $z^{[j]} \rightarrow \infty$, and $\vartheta_j(z_j) \rightarrow -\infty$ as $z^{[j]} \rightarrow -\infty$.
- Then $\Delta^{\text{ATE}}_{j,\mathcal{J}\setminus j}(x, z)$ and $\Delta^{\text{LATE}}_{j}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ are arbitrarily close when evaluated at a sufficiently large value of $\tilde{z}^{[j]}$ and a sufficiently small value of $z^{[j]}$.
- Following Heckman and Vytlacil (1999), $\Delta^{\text{TT}}_{j,\mathcal{J}\setminus j}(x, z)$ and $\Delta^{\text{LATE}}_{j}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ are arbitrarily close for sufficiently small $z^{[j]}$.

- Our discussion has focused on the Wald estimands.
- Alternatively we could also follow Heckman and Vytlacil (1999, 2001, 2005) in expressing ATE and TT as integrated versions of MTE.
- By theorem 3, we can use LIV to identify MTE and can thus express ATE and TT as integrated versions of the LIV estimand.

- For a general instrument $J\left(Z^{[j]}, Z^{[-j]}\right)$ constructed from $\left(Z^{[j]}, Z^{[-j]}\right)$, which we denote as $J^{[j]}$, we can obtain a parallel construction to the characterization of standard IV:

$$\Delta_{J^{[j]}}^{\mathrm{IV}} = \int_0^1 \Delta^{\mathrm{MTE}}\left(x, z, u_{D_j}\right) \omega_{\mathrm{IV}}^{J^{[j]}}\left(u_{D_j}\right) du_{D_j}, \qquad (50)$$

where

$$\omega_{\mathrm{IV}}^{J^{[j]}} = \frac{E\left[J^{[j]} - E\left(J^{[j]}\right) \mid P_j(Z) \geq u_{D_j}\right] \Pr\left(P_j(Z) \geq u_{D_j} \mid Z^{[-j]} = z^{[-j]}\right)}{\operatorname{Cov}(Z^{[j]}, D_{\mathcal{J},j})}, \qquad (51)$$

where u_{D_j} is defined at the beginning of this section and where we keep the conditioning on $X = x$ implicit.

- Note that from Theorem 3, we obtain that

$$\begin{aligned}\frac{\frac{\partial}{\partial z^{[j]}} E\left[Y \mid X=x, Z=z\right]}{\frac{\partial P_j(z)}{\partial z^{[j]}}} &= \frac{\partial E[Y \mid X=x, Z=z]}{\partial P_j(z)} \\ &= E\left[Y_j - Y_{\mathcal{J}\setminus j} \mid X=x, Z=z, \vartheta_j\left(Z_j\right) - V_j = R_{\mathcal{J}\setminus j}\left(Z\right)\right]\end{aligned}$$

  so we obtain that LIV identifies MTE and linear IV is a weighted average of LIV with the weights summing to one.
- These results mirror the results established in the binary case.

- In the literature on the effects of schooling ($S = \sum_{j \in \mathcal{J}} j D_{\mathcal{J},j}$) on earnings ($Y_{\mathcal{J}}$), it is conventional to instrument S.
- Our website presents an analysis of this case.
- For the general unordered case,

$$\Delta^{\text{IV}}_{J^{[j]}} = \frac{\text{Cov}(J^{[j]}, Y_{\mathcal{J}})}{\text{Cov}(J^{[j]}, S)}$$

  can be decomposed into economically interpretable components where the weights can be identified but the objects being weighted cannot be identified using local instrumental variables or LATE without making large support assumptions.
- However, the components can be identified using a structural model.

- The trick we have used in this section comparing outcomes in j to the next best option converts a general unordered multiple outcome model into a two outcome setup.
- This effectively partitions $Y_{\mathcal{J}}$ into two components, as in (48). Thus we write

$$Y_{\mathcal{J}} = D_{\mathcal{J},j} Y_j + (1 - D_{\mathcal{J},j})\, Y_{\mathcal{J} \setminus j},$$

where

$$Y_{\mathcal{J} \setminus j} = \sum_{\substack{\ell \neq j \\ \ell \in \mathcal{J}}} \frac{D_{\mathcal{J},\ell}}{1 - D_{\mathcal{J},j}} Y_\ell \times \mathbf{1}\left(D_{\mathcal{J},j} \neq 1\right).$$

In the more general unordered case with three or more choices, to analyze IV estimates of the effect of S on $Y_{\mathcal{J}}$, we must work with $Y_{\mathcal{J}} = \sum_{k \in \mathcal{J}} D_{\mathcal{J},k} Y_k$ and make multiple comparisons across potential outcomes.

- This requires us to move outside of the LATE/LIV framework, which is inherently based on binary comparisons.
- We consider models that do not impose additive separability.
- This includes a general random coefficient model.

- Comparing policy p to policy p',

$$E\left(Y_p \mid X\right) - E\left(Y_{p'} \mid X\right) = \int_0^1 E(\Delta \mid X, U_D = u_D)(F_{P_{p'}|X}(u_D) - F_{P_p|X}(u_D))\, du_D,$$

which gives the required weights.

- Recall $\Delta = Y_1 - Y_0$ and we can drop the p, p' subscripts on outcomes and errors.

Roy Model

$$
\begin{aligned}
Y_1 &= \mu_1 + U_1; \\
Y_0 &= \mu_0 + U_0; \\
I &= Z\gamma - V; \\
D &= \mathbf{1}\left[I > 0\right]
\end{aligned}
$$

Propensity Score

The propensity score conditional on Z:

$$D = \mathbf{1}[I > 0] = \mathbf{1}[Z\gamma > V]$$

The propensity score:

$$P(Z) \equiv E[D|Z] = \Pr(D = 1|Z) = \Pr(\gamma Z > V) = F_V(Z\gamma)$$

Definition:

$$F_V(V) \equiv U_D$$

therefore

$$\gamma Z > V \Leftrightarrow F_V(\gamma Z) > U_D \Leftrightarrow P(Z) > U_D$$

$$E[D] = \int_{-\infty}^{\infty} P(z) f_Z(z)\, dz$$

$$E(D) = E(E(\mathbf{1}[P(Z) > U_D] | U_D))$$

$$= 1 - E\left(F_{P(Z)}(U_D)\right)$$

$$F_{P(Z)}(p) = \Pr\left(Z < F_V^{-1}(p)\right) = F_Z\left(F_V^{-1}(p)\right)$$

The Normality Assumption

Normality assumptions

$$\begin{pmatrix} U_1 \\ U_0 \\ V \end{pmatrix} \sim N(0, \Sigma);\Sigma \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{10} & \sigma_{V1} \\ \cdot & \sigma_0^2 & \sigma_{V0} \\ \cdot & \cdot & \sigma_V^2 \end{pmatrix}$$

$$\Rightarrow \begin{bmatrix} U_1 - U_0 \\ V \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 + \sigma_0^2 - 2\sigma_{10} & \sigma_{1V} - \sigma_{0V} \\ \sigma_{1V} - \sigma_{0V} & \sigma_V^2 \end{bmatrix}\right)$$

The Propensity Score $P(Z)$

$$P(Z) = \Pr(\gamma Z > V) = \Phi\left(\frac{\gamma Z}{\sigma_V}\right)$$

Propensity Score under normality assumptions

$$F_{P(Z)}(t) = \Pr\left(F_V(Z) < t\right) = \Pr\left(Z < F_V^{-1}(t)\right) = F_{P(Z)}\left(F_V^{-1}(t)\right)$$

$$= \Phi\left(\frac{F_V^{-1}(t) - \mu_Z}{\sigma_Z}\right) = \Phi\left(\frac{\Phi^{-1}(t)\cdot\sigma_V - \mu_Z}{\sigma_Z}\right)$$

$$f_{P(Z)}(t) = \frac{\partial F_{P(Z)}(t)}{\partial t} = \phi\left(\frac{\Phi^{-1}(t)\cdot\sigma_V - \mu_Z}{\sigma_Z}\right)\frac{\sigma_V}{\sigma_Z}\cdot\frac{1}{\phi\left(\Phi^{-1}(t)\right)}$$

Marginal Treatment Effect (MTE) and Average Treatment Effect (ATE):

$$\begin{aligned} ATE &= E\left[Y_1 - Y_0\right] = \mu_1 - \mu_0 \\ MTE\left(v\right) &= E\left[Y_1 - Y_0 | V = v\right] \\ &= ATE + E\left[U_1 - U_0 | V = v\right] \end{aligned}$$

The MTE based on U_D:

$$\begin{aligned} MTE\left(u_D\right) &= E\left[Y_1 - Y_0 | U_D = u_D\right] \\ &= ATE - E\left[U_1 - U_0 | U_D = u_D\right] \end{aligned}$$

Whenever $U_D = P\left(Z\right)$ the agent is indifferent between treatments.

Under Normality Assumptions

$$\Rightarrow [U_1 - U_0 | V = v] \sim N\left(\frac{\sigma_{1-0,V}}{\sigma_V^2} \cdot v, \sigma^2\left(1-\rho^2\right)\right)$$

$$\Rightarrow MTE(v) = ATE + \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V} \cdot \frac{v}{\sigma_V}$$

Writing in terms of

$$\begin{aligned}
U_D &= F_V(V) = \Phi\left(\frac{V}{\sigma_V}\right) \Rightarrow V = \sigma_V \cdot \Phi^{-1}(U_D) \\
MTE(u_D) &= ATE + \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V^2} \cdot F_V^{-1}(u_D) \\
MTE(u_D) &= ATE + \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V} \cdot \Phi^{-1}(u_D)
\end{aligned}$$

Average Treatment Effect (ATE):

$$\begin{aligned} ATE &= E\left[E\left[Y_1 - Y_0 | V = v\right]\right] = \mu_1 - \mu_0 \\ &= E\left[E\left[MTE\left(v\right)|V = v\right]\right] \\ &= \int_{-\infty}^{\infty} MTE\left(v\right) \cdot \omega_{ATE}\left(v\right) f_v\left(v\right) dv \\ \omega_{ATE}\left(v\right) &= 1 \end{aligned}$$

Using U_D approach we obtain:

$$
\begin{aligned}
F_V(V) &\equiv U_D \\
ATE &= E\left[E\left[MTE(v) \,|\, U_D = u_D\right]\right] \\
ATE &= \int_0^1 MTE(u_D) \cdot \omega_{ATE}(u_D)\, du_D \\
\omega_{ATE}(u_D) &= 1
\end{aligned}
$$

The Treatment on the Treated

The relationship between the treatment on treated parameter and the marginal treatment effect is obtained below. First we do treatment on the treated given z.

$$\begin{aligned}
TT(z) &= E\left[Y_1 - Y_0 | I > 0, Z = z\right] = TT(P(Z)) \\
&= \frac{E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[I > 0\right], Z = z\right]}{\Pr\left(I > 0\right)} \\
&\quad \text{by law of iterated expectations} \\
&= \frac{E\left[\left(Y_1 - Y_0\right) \cdot \mathbf{1}\left[z\gamma > V\right]\right]}{\Pr\left(P\left(z\right) > U_D\right)} \\
&= \frac{\int_{-\infty}^{z\gamma} MTE\left(v\right) f_V(v) dv}{P\left(z\right)}
\end{aligned}$$

$$
\begin{aligned}
TT(P(Z)) &= E\left[Y_1 - Y_0 | I > 0\right] \\
&= \frac{E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[I > 0\right]\right]}{\Pr\left(I > 0\right)} \\
&\quad \text{by law of iterated expectations} \\
&= \frac{E\left[\left(Y_1 - Y_0\right) \cdot \mathbf{1}\left[P\left(Z\right) > U_D\right], Z = z\right]}{\Pr\left(P\left(Z\right) > U_D\right)} \\
&= \frac{\int_0^{P(z)} MTE\left(u_D\right) du_D}{P\left(z\right)}
\end{aligned}
$$

Using Normality Assuptions

$$
\begin{aligned}
TT(Z) &= E\left[Y_1 - Y_0 | I > 0, Z = z\right] \\
&= ATE + E\left[U_1 - U_0 | z\gamma > V, Z = z\right] \\
\text{define } \sigma &\equiv \sqrt{\sigma_1^2 + \sigma_0^2 - 2\sigma_{10}} \\
&= ATE + \sigma E\left[\frac{U_1 - U_0}{\sigma} | -\frac{V}{\sigma_V} > -\frac{z\gamma}{\sigma_V}\right] \\
&\Rightarrow \quad TT(z\gamma) = x\left(\beta_1 - \beta_0\right) - \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V} \cdot \lambda\left(-\frac{z\gamma}{\sigma_V}\right) \\
\text{Where } &: \\
\lambda(x) &\equiv \frac{\phi(x)}{1 - \Phi(x)} = \frac{\phi(x)}{\Phi(-x)}
\end{aligned}
$$

The propensity score is defined as $\Pr(D = 1|Z = z)$, where the conditional on Z is not used below in order to save notation. Based on the normality assumptions, we can obtain the following formulas:

$$P(z) = \Phi\left(\frac{z\gamma}{\sigma_V}\right) \text{ (Under Normality)}$$

Including this equation in the Treatment on treated effect we obtain:

$$\begin{aligned} TT(z) &= ATE - \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V} \cdot \lambda\left(-\frac{z\gamma}{\sigma_V}\right) \\ TT(P(z)) &= ATE - \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V} \cdot \frac{\phi\left(\Phi^{-1}(P(z))\right)}{P(z)} \end{aligned}$$

$$
\begin{aligned}
TT &= E\left[Y_1 - Y_0 | I > 0\right] \\
&= \frac{E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[I > 0\right]\right]}{\Pr\left(I > 0\right)} \\
&\quad \text{by law of iterated expectations} \\
&= \frac{E\left[E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[Z\gamma > v\right]\right] | V = v\right]}{\Pr\left(Z\gamma > V\right)} \\
&\quad \text{but } Y_1, Y_0 | V \perp\!\!\!\perp D | V,
\end{aligned}
$$

$$
\begin{aligned}
& \text{using Fubini's theorem} \\
&= \frac{E\left[E\left[Y_1 - Y_0 | V = v\right] \cdot E\left[\mathbf{1}\left[Z\gamma > v\right] | V = v\right]\right]}{\Pr\left(Z\gamma > V\right)} \\
&= E\left[MTE\left(v\right) \cdot \frac{E\left[\mathbf{1}\left[Z\gamma > v\right] | V = v\right]}{\Pr\left(Z\gamma > V\right)}\right] \\
&= \int_{-\infty}^{\infty} E\left[MTE\left(v\right) \cdot \omega_{TT}\left(v\right) f_v\left(v\right) dv\right] \\
\omega_{TT}\left(v\right) &= \frac{E\left[\mathbf{1}\left[Z\gamma > v\right] | V = v\right]}{\Pr\left(Z\gamma > V\right)} = \frac{1 - F_{Z\gamma}\left(v\right)}{E\left(D\right)}
\end{aligned}
$$

The same analysis using the propensity score:

$$
\begin{aligned}
TT &= E\left[Y_1 - Y_0 | I > 0\right] \\
&= \frac{E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[I > 0\right]\right]}{\Pr\left(I > 0\right)} \\
&\quad \text{by law of iterated expectations} \\
&= \frac{E\left[E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[P\left(Z\right) > u_D\right]\right] | U_D = u_D\right]}{\Pr\left(P\left(Z\right) > U_D\right)}; U_D \equiv F_V\left(V\right) \\
&\quad \text{but } Y_1, Y_0 | U_D \perp\!\!\!\perp D | U_D,
\end{aligned}
$$

using Fubini's theorem

$$
\begin{aligned}
&= \frac{E\left[E\left[Y_1 - Y_0 | U_D = u_D\right] \cdot E\left[\mathbf{1}\left[P(Z) > u_D\right] | U_D = u_D\right]\right]}{E(P(Z))} \\
&= E\left[MTE(u_D) \cdot \frac{E\left[\mathbf{1}\left[P(Z) > u_D\right] | U_D = u_D\right]}{E(P(Z))}\right] \\
&= \int_{-\infty}^{\infty} MTE(u_D) \cdot \omega_{TT}(u_D)\, du_D
\end{aligned}
$$

$$
\begin{aligned}
\text{Observe that } U_D &\sim Uniform\,[0,1] \\
\omega_{TT}(u_D) &= \frac{E\left[\mathbf{1}\left[P(Z) > u_D\right] | U_D = u_D\right]}{E(P(Z))} \\
&= \frac{\int_{u_D}^{1} f_{P(Z)}(p)\, dp}{E(P(Z))} = \frac{1 - F_{P(Z)}(u_D)}{E(P(Z))}
\end{aligned}
$$

The Treatment on the Untreated

The relationship between the treatment on untreated parameter and the marginal treatment effect is obtained below:

$$
\begin{aligned}
TUT &= E\left[Y_1 - Y_0 | I \leqslant 0, Z = z\right] \\
&= \frac{E\left[\left(Y_1 - Y_0\right) \cdot \mathbf{1}\left[I \leqslant 0\right], Z = z\right]}{\Pr\left(I \leqslant 0\right)} \\
&\quad \text{by law of iterated expectations} \\
&= \frac{E\left[E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[z\gamma \leqslant v\right]\right] | V = v\right]}{\Pr\left(z\gamma \leqslant V\right)} \\
&\quad \text{but } Y_1, Y_0 | V \perp\!\!\!\perp D | V,
\end{aligned}
$$

$$
\begin{aligned}
& \text{using Fubini's theorem} \\
= \quad & \frac{E\left[E\left[Y_1 - Y_0 | V = v\right] \cdot E\left[\mathbf{1}\left[z\gamma \leqslant v\right] | V = v\right]\right]}{\Pr\left(z\gamma \leqslant V\right)} \\
= \quad & E\left[MTE\left(v\right) \cdot \frac{E\left[\mathbf{1}\left[z\gamma \leqslant v\right] | V = v\right]}{\Pr\left(z\gamma \leqslant V\right)}\right] \\
= \quad & \int_{-\infty}^{\infty} MTE\left(v\right) \cdot \omega_{TUT}\left(v\right) f_v\left(v\right) dv
\end{aligned}
$$

$$
\begin{aligned}
\omega_{TUT}\left(v\right) \quad = \quad & \frac{E\left[\mathbf{1}\left[z\gamma \leqslant v\right] | V = v\right]}{\Pr\left(z\gamma \leqslant V\right)} = \frac{E\left[\mathbf{1}\left[z\gamma \leqslant v\right] | V = v\right]}{1 - \Pr\left(z\gamma > v\right)} \\
= \quad & \frac{\int_{-\infty}^{v} f_{z\gamma}\left(z\right) dz}{1 - \Pr\left(z\gamma > V\right)} = \frac{F_{z\gamma}\left(v\right)}{1 - E\left(D\right)}
\end{aligned}
$$

The same analysis can be done with the propensity score approach:

$$
\begin{aligned}
TUT &= E\left[Y_1 - Y_0 | I \leqslant 0\right] \\
&= \frac{E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[I \leqslant 0\right]\right]}{\Pr\left(I \leqslant 0\right)} \\
&\quad \text{by law of iterated expectations} \\
&= \frac{E\left[E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[P\left(Z\right) \leqslant u_D\right]\right] | U_D = u_D\right]}{\Pr\left(P\left(Z\right) \leqslant U_D\right)} \\
U_D &\equiv F_V\left(V\right) \\
&\quad \text{but } Y_1, Y_0 | U_D \perp\!\!\!\perp D | U_D,
\end{aligned}
$$

using the Fubini's theorem

$$= \frac{E\left[E\left[Y_1 - Y_0 | U_D = u_D\right] \cdot E\left[\mathbf{1}\left[P(Z) \leqslant u_D\right] | U_D = u_D\right]\right]}{1 - E(P(Z))}$$

$$= E\left[MTE(u_D) \cdot \frac{E\left[\mathbf{1}\left[P(Z) \leqslant u_D\right] | U_D = u_D\right]}{1 - E(P(Z))}\right]$$

$$\begin{aligned}
\text{Observe that } U_D &\sim Uniform[0,1] \\
&= \int_{-\infty}^{\infty} E\left[MTE(u_D) \cdot \omega_{TUT}(u_D)\, du_D\right] \\
\omega_{TUT}(u_D) &= \frac{E\left[\mathbf{1}\left[P(Z) \leqslant u_D\right] | U_D = u_D\right]}{1 - E(P(Z))} \\
&= \frac{\int_0^{u_D} f_{P(Z)}(p)\, dp}{1 - E(P(Z))} = \frac{F_{P(Z)}(u_D)}{1 - E(P(Z))}
\end{aligned}$$

$$
\begin{aligned}
TUT(Z) &= E\left[Y_1 - Y_0 | I < 0\right] \\
&= \frac{E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[I < 0\right]\right]}{\Pr\left(I < 0\right)} \\
&\qquad \text{by law of iterated expectations} \\
&= \frac{E\left[\left(Y_1 - Y_0\right) \cdot \mathbf{1}\left[\gamma Z < V\right]\right]}{\Pr\left(P\left(Z\right) < U_D\right)} \\
&= \frac{\int\limits_{\gamma Z}^{\infty} MTE\left(v\right) f_V\left(v\right) dv}{1 - P\left(Z\right)}
\end{aligned}
$$

$$
\begin{aligned}
TUT(P(Z)) &= E\left[Y_1 - Y_0 | I < 0\right] \\
&= \frac{E\left[Y_1 - Y_0 \cdot \mathbf{1}\left[I < 0\right]\right]}{\Pr\left(I < 0\right)} \\
&\quad \text{by law of iterated expectations} \\
&= \frac{E\left[\left(Y_1 - Y_0\right) \cdot \mathbf{1}\left[P\left(Z\right) < U_D\right]\right]}{\Pr\left(P\left(Z\right) < U_D\right)} \\
&= \frac{\int_{P(Z)}^{1} MTE\left(u_D\right) du_D}{1 - P\left(Z\right)}
\end{aligned}
$$

Using Normality Assumptions

$$
\begin{aligned}
TUT(Z\gamma) &= E\left[Y_1 - Y_0 | I \leqslant 0\right] \\
&= \alpha_1 - \alpha_0 + X\left(\beta_1 - \beta_0\right) + E\left[U_1 - U_0 | Z\gamma \leqslant V\right] \\
&= ATE + E\left[U_1 - U_0 | Z\gamma \leqslant V\right]
\end{aligned}
$$

$$
\begin{aligned}
\text{define } \sigma &= \sqrt{\sigma_1^2 + \sigma_0^2 - 2\sigma_{10}}, \lambda\left(x\right) \equiv \frac{\phi\left(x\right)}{\Phi(-x)} \\
&= ATE + \sigma E\left[\frac{U_1 - U_0}{\sigma} | \frac{V}{\sigma_V} \geqslant \frac{Z\gamma}{\sigma_V}\right] \\
&\Rightarrow TUT(Z\gamma) = X\left(\beta_1 - \beta_0\right) + \frac{\sigma_{1V} - \sigma_{0V}}{\sigma_V} \cdot \lambda\left(\frac{Z\gamma}{\sigma_V}\right)
\end{aligned}
$$

OLS (Matching)

The relationship between the OLS parameter and the marginal treatment effect is obtained below:

$$
\begin{aligned}
\Delta_{matching} &= E\left[Y_1|D=1\right] - E\left[Y_0|D=0\right] \\
&= ATE + E\left[U_1|Z\gamma > V\right] - E\left[U_0|Z\gamma \leqslant V\right] \\
&= ATE + \frac{E\left[U_1 \cdot \mathbf{1}\left[Z\gamma > V\right]\right]}{\Pr\left(Z\gamma > V\right)} - \frac{E\left[U_0 \cdot \mathbf{1}\left[Z\gamma \leqslant V\right]\right]}{\Pr\left(Z\gamma \leqslant V\right)} \\
&= ATE + E\left[\begin{array}{c} \frac{E[U_1 \cdot \mathbf{1}[Z\gamma > v]|V=v]}{\Pr(Z\gamma > V)} \\ -\frac{E[U_0 \cdot \mathbf{1}[Z\gamma \leqslant v]|V=v]}{\Pr(Z\gamma \leqslant V)} \end{array}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\begin{array}{c} ATE\left(v\right) + \frac{E[U_1 \cdot \mathbf{1}[Z\gamma > v] | V = v]}{\Pr(Z\gamma > V)} \\ - \frac{E[U_0 \cdot \mathbf{1}[Z\gamma \leqslant v] | V = v]}{\Pr(Z\gamma \leqslant V)} \end{array}\right] \\
&= E\left[MTE\left(v\right) \cdot \left(\begin{array}{c} \omega_{ATE}\left(v\right) + \frac{E[U_1 \cdot \mathbf{1}[Z\gamma > v] | V = v]}{MTE(v) \cdot \Pr(Z\gamma > V)} - \\ \frac{E[U_0 \cdot \mathbf{1}[Z\gamma \leqslant v] | V = v]}{MTE(v) \cdot \Pr(Z\gamma \leqslant V)} \end{array}\right)\right] \\
&= E\left[MTE\left(v\right) \cdot \left(\begin{array}{c} 1 + \frac{E[U_1 \cdot \mathbf{1}[Z\gamma > v] | V = v]}{MTE(v) \cdot \Pr(Z\gamma > V)} - \\ \frac{E[U_0 \cdot \mathbf{1}[Z\gamma \leqslant v] | V = v]}{MTE(v) \cdot \Pr(Z\gamma \leqslant V)} \end{array}\right)\right] \\
&= E\left[MTE\left(V\right) \cdot \omega_{match}\left(V\right)\right] = \int\limits_{-\infty}^{\infty} MTE\left(v\right) \cdot \omega_{match}\left(v\right) f_v\left(v\right) dv
\end{aligned}
$$

$$\omega_{match}(v) = 1 + \frac{E[U_1 \cdot \mathbf{1}[Z\gamma > v] | V = v]}{MTE(v) \cdot \Pr(Z\gamma > V)} - \frac{E[U_0 \cdot \mathbf{1}[Z\gamma \leqslant v] | V = v]}{MTE(v) \cdot \Pr(Z\gamma \leqslant V)}$$

$$U_1, U_0 | V \perp\!\!\!\perp Z$$

$$\begin{aligned} E[U_1 \cdot \mathbf{1}[Z\gamma > v] | V = v] &= E[U_1 | V = v] \cdot (1 - F_{Z\gamma}(v)) \\ E[U_0 \cdot \mathbf{1}[Z\gamma \leqslant v] | V = v] &= E[U_0 | V = v] \cdot F_{Z\gamma}(v) \end{aligned}$$

$$\omega_{match}(v) = 1 + \frac{E[U_1 | V = v] \cdot (1 - F_{Z\gamma}(v))}{MTE(v) \cdot \Pr(Z\gamma > V)} - \frac{E[U_0 | V = v] \cdot F_{Z\gamma}(v)}{MTE(v) \cdot \Pr(Z\gamma \leqslant V)}$$

The same analysis can be done with the propensity score:

$$
\begin{aligned}
\Delta_{matching} &= E\left[Y_1|D=1\right] - E\left[Y_0|D=0\right] \\
&= ATE + E\left[U_1|P(Z) > U_D\right] - E\left[U_0|P(Z) \leqslant U_D\right] \\
&= E\left[\begin{array}{c} ATE(u_D) + \frac{E[U_1 \cdot \mathbf{1}[P(Z)>u_D]|U_D=u_D]}{\Pr(P(Z)>U_D)} \\ -\frac{E[U_0 \cdot \mathbf{1}[P(Z)\leqslant u_D]|U_D=u_D]}{\Pr(P(Z)\leqslant U_D)} \end{array}\right] \\
&= E\left[MTE(u_D) \cdot \left(\begin{array}{c} 1 + \frac{E[U_1 \cdot \mathbf{1}[P(Z)>u_D]|U_D=u_D]}{MTE(u_D)\cdot\Pr(P(Z)>U_D)} - \\ -\frac{E[U_0 \cdot \mathbf{1}[P(Z)\leqslant u_D]|U_D=u_D]}{MTE(u_D)\cdot\Pr(P(Z)\leqslant U_D)} \end{array}\right)\right] \\
&= E\left[MTE(u_D) \cdot \omega_{OLS}(u_D)\right] \\
&= \int_{-\infty}^{\infty} MTE(u_D) \cdot \omega_{OLS}(u_D)\, du_D
\end{aligned}
$$

$$\omega_{match}(u_D) = 1 + \frac{E[U_1 \cdot \mathbf{1}[P(Z) > u_D] | U_D = u_D]}{MTE(u_D) \cdot \Pr(P(Z) > U_D)} - \frac{E[U_0 \cdot \mathbf{1}[P(Z) \leqslant u_D] | U_D = u_D]}{MTE(u_D) \cdot \Pr(P(Z) \leqslant U_D)}$$

Using Normality Assumption

$$\begin{aligned}\omega_{match}(u_D) = & \quad 1 + \frac{E[U_1 \cdot \mathbf{1}[Z\gamma > v] | V = v]}{MTE(v) \cdot \Pr(Z\gamma > V)} - \frac{E[U_0 \cdot \mathbf{1}[Z\gamma \leqslant v] | V = v]}{MTE(v) \cdot \Pr(Z\gamma \leqslant V)} \\ = & \quad 1 + \frac{E[U_1 | V = v] \cdot E[\mathbf{1}[Z\gamma > v]]}{MTE(v) \cdot \Pr(Z\gamma > V)} - \frac{E[U_0 \cdot | V = v] \cdot E[\mathbf{1}[Z\gamma \leqslant v]]}{MTE(v) \cdot \Pr(Z\gamma \leqslant V)}\end{aligned}$$

$$= \quad 1 + \frac{\left(\frac{\sigma_{1V}}{\sigma_V^2} \cdot v\right) \cdot \Phi\left(\frac{\gamma \cdot \mu_Z - v}{\sqrt{\gamma' \boldsymbol{\Sigma} \gamma}}\right)}{MTE(v) \cdot \Phi\left(\frac{\gamma \cdot \mu_Z}{\sqrt{\gamma' \boldsymbol{\Sigma} \gamma + \sigma_V}}\right)} - \frac{\left(\frac{\sigma_{0V}}{\sigma_V^2} \cdot v\right) \cdot \Phi\left(\frac{v - \gamma \cdot \mu_Z}{\sqrt{\gamma' \boldsymbol{\Sigma}_Z \gamma}}\right)}{MTE(v) \cdot \Phi\left(-\frac{\gamma \cdot \mu_Z}{\sqrt{\gamma' \boldsymbol{\Sigma} \gamma + \sigma_V}}\right)}$$

Matching in Z using normality assumptions

$$\Delta_{matching} \quad = \quad E\left(Y_1 | D = 1\right) - E\left(Y_0 | D = 0\right)$$

Matching in Z:

$$
\begin{aligned}
&= ATE + E\left(U_1 | Z\gamma' > V\right) - E\left(U_0 | Z\gamma' < V\right) \\
&= ATE + E\left(U_1 | -V > -Z\gamma'\right) - E\left(U_0 | V > Z\gamma'\right) \\
&= ATE + E\left(U_1 | -\frac{V}{\sigma_V} > -\frac{Z\gamma'}{\sigma_V}\right) - E\left(U_0 | \frac{V}{\sigma_V} > \frac{Z\gamma'}{\sigma_V}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= ATE + \sigma_1 E\left(\frac{U_1}{\sigma_1} | -\frac{V}{\sigma_V} > -\frac{Z\gamma'}{\sigma_V}\right) - \sigma_0 E\left(\frac{U_0}{\sigma_0} | \frac{V}{\sigma_V} > \frac{Z\gamma'}{\sigma_V}\right) \\
&= ATE - \frac{\sigma_{1V}}{\sigma_V} \cdot \lambda\left(-\frac{\gamma Z}{\sigma_V}\right) - \frac{\sigma_{0V}}{\sigma_V} \cdot \lambda\left(\frac{\gamma Z}{\sigma_V}\right) \\
&= ATE - \left(\frac{\frac{\sigma_{1V}}{\sigma_V} \cdot \Phi\left(-\frac{Z \cdot \gamma'}{\sigma_V}\right) + \frac{\sigma_{0V}}{\sigma_V} \cdot \Phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right)}{\Phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right) \Phi\left(-\frac{Z \cdot \gamma'}{\sigma_V}\right)}\right) \phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right)
\end{aligned}
$$

Matching in $P(Z)$ using normality assumptions

$$\Delta_{matching} = E(Y_1|D=1) - E(Y_0|D=0)$$

Matching in $P(Z)$:

$$\begin{aligned}
&= ATE + E(U_1|Z\gamma' > V) - E(U_0|Z\gamma' < V) \\
&= ATE - \frac{\sigma_{1V}}{\sigma_V} \cdot \lambda\left(-\frac{\gamma Z}{\sigma_V}\right) - \frac{\sigma_{0V}}{\sigma_V} \cdot \lambda\left(\frac{\gamma Z}{\sigma_V}\right) \\
&= ATE - \left(\frac{\sigma_{1V}}{\sigma_V} \cdot \frac{1}{P(Z)} + \frac{\sigma_{0V}}{\sigma_V} \cdot \frac{1}{1-P(Z)}\right) \phi\left(\Phi^{-1}(P(Z))\right) \\
&= ATE - \left(\frac{\frac{\sigma_{1V}}{\sigma_V} \cdot (1-P(Z)) + \frac{\sigma_{0V}}{\sigma_V} \cdot P(Z)}{P(Z)(1-P(Z))}\right) \phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right)
\end{aligned}$$

The PRTE

$$
\begin{aligned}
& E\left(Y_1 - Y_0 | P(Z) - U_D = t\right) \\
= \; & E\left(Y_1 - Y_0 | F_V\left(Z\right) - U_D = t\right) \\
= \; & E\left(E\left(Y_1 - Y_0 | F_V\left(Z\right) = p, p - U_D = t\right) | F_V\left(Z\right) - U_D = t\right) \\
= \; & E\left(E\left(Y_1 - Y_0 | U_D = p - t\right) | F_V\left(Z\right) - U_D = t\right) \\
= \; & E\left[MTE(p - t) | P(Z) - U_D = t\right] \\
= \; & \int_0^1 MTE(p - t) f_P(p) dp = \int_0^1 MTE(p) f_P(p + t) dp \\
\upsilon \notin \; & [0, 1] \Rightarrow f_P(\upsilon) = MTE(\upsilon) = 0
\end{aligned}
$$

$$
\begin{aligned}
& E\left(Y_1 - Y_0 \middle| -t < P(Z) - U_D < t\right) \\
= \; & E\left(E\left(Y_1 - Y_0 \middle| P(Z) - U_D = \xi\right) \middle| -t < P(Z) - U_D < t\right) \\
\Theta \equiv \; & P(Z) - U_D
\end{aligned}
$$

$$
\begin{aligned}
f_\Theta(\theta) &= \int f_{P(Z)}(\theta) \cdot f_{U_D}(\theta) \\
&= E\left(E\left(Y_1 - Y_0 \middle| \Theta = \xi\right) \middle| -t < \Theta < t\right) \\
&= \frac{E\left(E\left(Y_1 - Y_0 \middle| \Theta = \xi\right) \cdot \mathbf{1}\left[-t < \Theta < t\right]\right)}{\Pr\left(-t < \Theta < t\right)} \\
&= \frac{E\left(\int\limits_{-t}^{t} E\left(Y_1 - Y_0 \middle| \Theta = \xi\right) F_{P(Z)}(\xi + 1)\, d\xi\right)}{\Pr\left(-t < \Theta < t\right)}
\end{aligned}
$$

$$= \frac{E\left(\left(\int_0^1 MTE(p) f_P(p+\xi)dp\right)\cdot \mathbf{1}\left[-t < P(Z) - U_D < t\right]\right)}{\Pr\left(-t < \Theta < t\right)}$$

$$= \frac{\int_{-t}^{t}\left(\int_0^1 MTE(p) f_P(p+\xi)dp\right) f_{P(Z)}\left(\xi + u_D\right) d\xi}{\Pr\left(-t < \Theta < t\right)}$$

$$= \frac{E\left(\left(\int_0^1 MTE(p) f_P(p+\xi)dp\right)\cdot \mathbf{1}\left[-t < P(Z) - U_D < t\right]\right)}{\Pr\left(-t < \Theta < t\right)}$$

$$= \frac{\int_{-t}^{t}\int_0^1 MTE(u_D) f_P(u_D + t^*) du_D dt^*}{\Pr\left(-t < P(Z) - U_D < t\right)}$$

$$
\begin{aligned}
\Pr(-t < \Theta < t) &= \Pr(-t < P(Z) - U_D < t) \\
&= E(\mathbf{1}[-t < P(Z) - U_D < t]) \\
&= E(E(\mathbf{1}[u_D - t < P(Z) < t + u_D] \,|\, U_D = u_D)) \\
&= E\left(F_{P(Z)}(t + U_D) - F_{P(Z)}(-t + U_D)\right) \\
&= \int_0^1 \left[F_{P(Z)}(t + u_D) - F_{P(Z)}(-t + u_D)\right] du_D \\
F_{P(Z)}(p) &= \Phi\left(\frac{\Phi^{-1}(p) \cdot \sigma_V - \mu_Z}{\sigma_Z}\right)
\end{aligned}
$$

$$
\begin{aligned}
&E(Y_1 - Y_0 | Z - V = t) \\
&= \int_0^1 MTE(u_D) \frac{f_Z(F_V^{-1}(u_D) + t)}{E(f_V(Z - t))} du_D
\end{aligned}
$$

therefore

$$
\begin{aligned}
& E\left(Y_1 - Y_0 \middle| -t < Z - V < t\right) \\
= \ & E\left(E\left(Y_1 - Y_0 \middle| Z - V = t\right) \middle| -t < Z - V < t\right) \\
= \ & \frac{E\left(E\left(Y_1 - Y_0 | Z - V = t\right) \cdot \mathbf{1}\left[-t < Z - V < t\right]\right)}{\Pr\left(-t < Z - V < t\right)} \\
= \ & \frac{\int_{-t}^{t} \int_{0}^{1} MTE(u_D) \frac{f_Z(F_V^{-1}(u_D) + t^*)}{E(f_V(Z - t^*))} du_D dt^*}{\Pr\left(-t < Z - V < t\right)}
\end{aligned}
$$

$$
\begin{aligned}
& \Pr\left(-t < Z - V < t\right) \\
= & \int_{-\infty}^{\infty} \left[F_Z\left(t + v\right) - F_Z\left(-t + v\right)\right] f_V\left(v\right) dv
\end{aligned}
$$

$$
\begin{aligned}
F_Z\left(z\right) &= \Phi\left(\frac{z - \mu_Z}{\sigma_Z}\right) \\
f_V\left(v\right) &= \phi\left(\frac{v}{\sigma_V}\right) \frac{1}{\sigma_V}
\end{aligned}
$$

$$
\begin{aligned}
& E\left(Y_1 - Y_0 | P(Z)/U_D = 1 - t\right) \\
= & \int_0^1 MTE(u_D) \frac{f_P(u_D/\left(1-t\right))\left(1-t\right)^2 u_D}{E\left(D\right)} du_D
\end{aligned}
$$

therefore

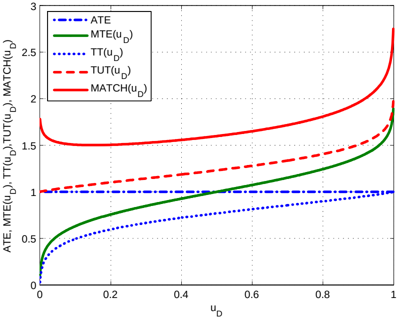
$$
\begin{aligned}
& E\left(Y_1 - Y_0 | 1 - t < P(Z)/U_D < 1 + t\right) \\
= \quad & E\left(E\left(Y_1 - Y_0 | P(Z)/U_D - 1 = -t^*\right) | 1 - t < P(Z)/U_D < 1 + t\right) \\
= \quad & \frac{E\left(E\left((Y_1 - Y_0 | P(Z)/U_D - 1 = -t^*\right) \cdot \mathbf{1}\left[-t < P(Z)/U_D - 1 < t\right]\right))}{\Pr\left(1 - t < P(Z)/U_D < 1 + t\right)} \\
= \quad & \frac{E\left(\left(\int\limits_0^1 MTE(u_D) \frac{f_P(u_D/(1-t^*))(1-t^*)^2 u_D}{E(D)} du_D\right) \cdot \mathbf{1}\left[-t < P(Z)/U_D - 1 < t\right]\right)}{\Pr\left(1 - t < P(Z)/U_D < 1 + t\right)} \\
= \quad & \frac{\int\limits_{1-t}^{1+t} \int\limits_0^1 MTE(u_D) \frac{f_P(u_D/(1-t^*))(1-t^*)^2 u_D}{E(D)} du_D dt^*}{\Pr\left(1 - t < P(Z)/U_D < 1 + t\right)}
\end{aligned}
$$

$$\begin{aligned}
& \Pr\left(1-t<P(Z)/U_D<1+t\right) \\
= \; & E\left(\mathbf{1}\left[1-t<P(Z)/U_D<1+t\right]\right) \\
= \; & E\left(E\left(\mathbf{1}\left[(1-t)\,u_D<P(Z)<(1+t)\,u_D\right] \mid U_D=u_D\right)\right) \\
= \; & E\left(\left[F_{P(Z)}\left((1+t)\cdot U_D\right)-F_{P(Z)}\left((1-t)\cdot U_D\right)\right]\right) \\
= \; & \int_0^1 \left[F_{P(Z)}\left((1+t)\cdot u_D\right)-F_{P(Z)}\left((1-t)\cdot u_D\right)\right] du_D
\end{aligned}$$

$$F_{P(Z)}\left(p\right) \;=\; \Phi\left(\frac{\Phi^{-1}\left(p\right)\cdot\sigma_V-\mu_Z}{\sigma_Z}\right)$$

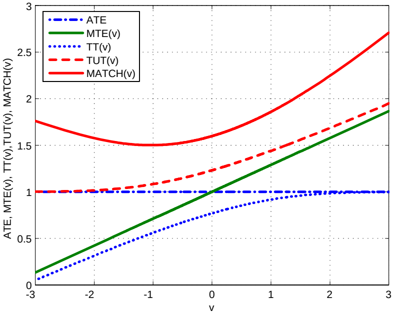
Treatment Effects in (u_D)

Figure A



Treatment Effects in (v)

Figure B



$$Y_1 = \alpha_1 + U_1; Y_0 = \alpha_0 + U_0 \qquad Z \perp\!\!\!\perp U_1, U_0, V$$

$$I = Z - V; D = \mathbf{1}[I > 0] = \mathbf{1}[Z > V]$$

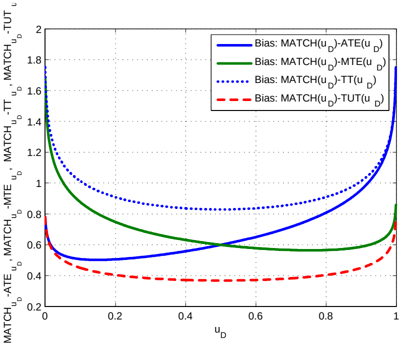
$$Y = DY_1 + (1 - D) Y_0 \qquad (U_1, U_0, V) \sim N(\mathbf{0}, \mathbf{\Sigma}_{U,V});$$

$$\Sigma_{U1,U0,V} \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{V1} & \sigma_{V0} \\ \cdot & \sigma_0^2 & \sigma_{10} \\ \cdot & \cdot & \sigma_V^2 \end{pmatrix} = \begin{pmatrix} 1.26 & 0.51 & -0.40 \\ \cdot & 2.01 & -0.90 \\ \cdot & \cdot & 3 \end{pmatrix}$$

$\mu_1 = 1; \mu_0 = 0;$

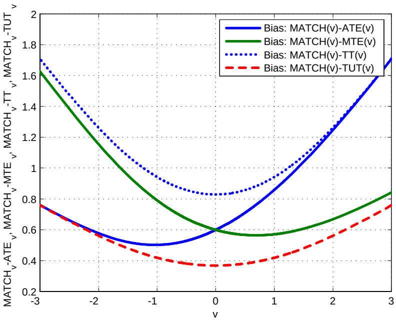
Treatment Effects Bias in (u_D)

Figure A



Treatment Effects Bias in (v)

Figure B



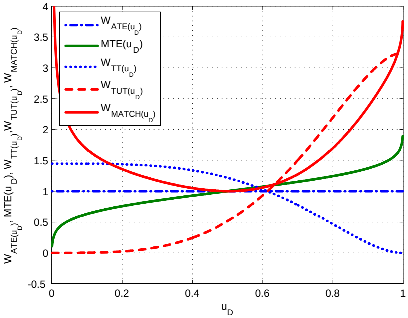
$$
\begin{array}{cc}
Y_1 = \alpha_1 + U_1; Y_0 = \alpha_0 + U_0 & Z \perp\!\!\!\perp U_1, U_0, V \\
I = Z - V; D = \mathbf{1}[I > 0] = \mathbf{1}[Z > V] & Z \sim N(\mu_Z, \sigma_Z^2) = N(1,1) \\
Y = DY_1 + (1 - D)Y_0 & (U_1, U_0, V) \sim N(\mathbf{0}, \mathbf{\Sigma}_{U,V});
\end{array}
$$

$$
\Sigma_{U1,U0,V} \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{V1} & \sigma_{V0} \\ \cdot & \sigma_0^2 & \sigma_{10} \\ \cdot & \cdot & \sigma_V^2 \end{pmatrix} = \begin{pmatrix} 1.26 & 0.51 & -0.40 \\ \cdot & 2.01 & -0.90 \\ \cdot & \cdot & 3 \end{pmatrix}
$$

$\mu_1 = 1; \mu_0 = 0;$

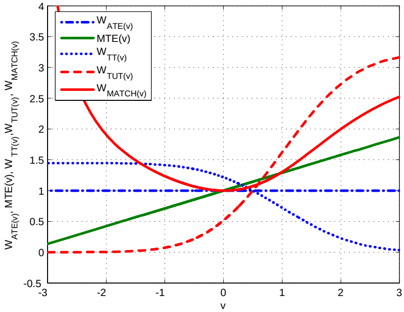
Treatment Weights (u_D)

Figure A



Treatment Effects Bias in (v)

Figure B



$$Y_1 = \alpha_1 + U_1; Y_0 = \alpha_0 + U_0 \qquad Z \perp\!\!\!\perp U_1, U_0, V$$
$$I = Z - V; D = \mathbf{1}[I > 0] = \mathbf{1}[Z > V] \quad Z \sim N(\mu_Z, \sigma_Z^2) = N(1,1)$$
$$Y = DY_1 + (1-D)Y_0 \qquad (U_1, U_0, V) \sim N(\mathbf{0}, \mathbf{\Sigma}_{U,V});$$
$$\Sigma_{U1,U0,V} \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{V1} & \sigma_{V0} \\ \cdot & \sigma_0^2 & \sigma_{10} \\ \cdot & \cdot & \sigma_V^2 \end{pmatrix} = \begin{pmatrix} 1.26 & 0.51 & -0.40 \\ \cdot & 2.01 & -0.90 \\ \cdot & \cdot & 3 \end{pmatrix}$$

$\mu_1 = 1; \mu_0 = 0;$

The Model

$$
\begin{aligned}
Y_1 &= \mu_1 + U_1; \\
Y_0 &= \mu_0 + U_0; \\
I &= Z \cdot \gamma' - V; \\
D &= \mathbf{1}\left[I > 0\right]
\end{aligned}
$$

$$
\Sigma_{U1,U0,V} \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{V1} & \sigma_{V0} \\ \cdot & \sigma_0^2 & \sigma_{10} \\ \cdot & \cdot & \sigma_V^2 \end{pmatrix}
$$

$$
\begin{bmatrix} U_1 - U_0 \\ V \end{bmatrix} \sim N\begin{pmatrix} \mathbf{0}, & \sigma_{1-0}^2 & \sigma_{V1} - \sigma_{V0} \\ & \cdot & \sigma_V^2 \end{pmatrix}
$$

$$
\sigma_{1-0} = \sqrt{\sigma_{U1}^2 + \sigma_{U0}^2 - 2\sigma_{10}}
$$

Propensity score:

$$\begin{aligned} P(Z) &\equiv \Pr(D = 1|Z) = P\left(\frac{Z \cdot \gamma'}{\sigma_V} > \frac{V}{\sigma_V}\right) \\ &= \Phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right) \end{aligned}$$

The transformation of variables:

$$\begin{aligned} P(Z) &= \Phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right) \Rightarrow \frac{Z \cdot \gamma'}{\sigma_V} = \Phi^{-1}(P(Z)) \\ 1 - P(Z) &= \Phi\left(-\frac{Z \cdot \gamma'}{\sigma_V}\right) \Rightarrow -\frac{Z \cdot \gamma'}{\sigma_V} = \Phi^{-1}(1 - P(Z)) \\ \Phi(\cdot) &\equiv \text{Standard Normal Probability Function.} \end{aligned}$$

Definitions:

$$\begin{aligned}\lambda(x) &= \frac{\phi(x)}{1-\Phi(x)} = \frac{\phi(x)}{\Phi(-x)}; \phi(x) = \frac{\partial \Phi(x)}{\partial x} \\ \lambda(x) &= E(X \mid X > x); X \sim N(0,1)\end{aligned}$$

Observe that:

$$
\begin{aligned}
\lambda\left(-\frac{Z\cdot\gamma'}{\sigma_V}\right) &= \frac{\phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right)}{\Phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right)} \\
\phi\left(\Phi^{-1}\left(1-P(Z)\right)\right) &= \phi\left(-\frac{Z\cdot\gamma'}{\sigma_V}\right) = \phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right) \\
&= \phi\left(\Phi^{-1}\left(P(Z)\right)\right) \\
\Phi\left(-\Phi^{-1}\left(P(Z)\right)\right) &= \Phi\left(-\frac{Z\cdot\gamma'}{\sigma_V}\right) = 1-\Phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right) \\
&= 1-\Phi\left(\Phi^{-1}\left(P(Z)\right)\right) \\
&= 1-P(Z) \\
\Phi\left(-\Phi^{-1}\left(1-P(Z)\right)\right) &= \Phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right) = \Phi\left(\Phi^{-1}\left(P(Z)\right)\right) = P(Z)
\end{aligned}
$$

The Ratio :

$$\lambda\left(\Phi^{-1}\left(P(Z)\right)\right) = \frac{\phi\left(\Phi^{-1}\left(P(Z)\right)\right)}{1-P(Z)}$$

$$\lambda\left(\Phi^{-1}\left(1-P(Z)\right)\right) = \frac{\phi\left(\Phi^{-1}\left(P(Z)\right)\right)}{P(Z)}$$

Treatment parameters :

$$
\begin{aligned}
ATE &\equiv E(Y_1 - Y_0) = \mu_1 - \mu_0 \\
\textbf{MTE in } V &= v\textbf{:} \\
MTE(v) &\equiv E(Y_1 - Y_0 | V = v) \\
&= ATE + E\left(U_1 - U_0 | \frac{V}{\sigma_V} = \frac{v}{\sigma_V}\right) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | \frac{V}{\sigma_V} = \frac{v}{\sigma_V}\right) \\
&= ATE + \frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V} \cdot \frac{v}{\sigma_V} \\
\text{If } v &= Z \cdot \gamma' \Rightarrow I = Z \cdot \gamma' - V = 0
\end{aligned}
$$

There is economic intition.

$$
\begin{aligned}
\textbf{MTE in } F_V(V) &= p\textbf{:} \\
MTE(p) &\equiv E\left(Y_1 - Y_0 | F_V(V) = p\right) \\
&= ATE + E\left(U_1 - U_0 | \frac{V}{\sigma_V} = \Phi^{-1}(p)\right) \\
&= ATE + \frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V} \cdot \Phi^{-1}(p) \\
\text{If } p &= F_V(Z \cdot \gamma') \Rightarrow I = F_V^{-1}(p) - V = 0 \\
& \text{There is economic intition.}
\end{aligned}
$$

Treatment parameters:

$$
\begin{aligned}
TT \text{ in } Z \quad &: \\
TT(Z) &\equiv E(Y_1 - Y_0 | D = 1, Z) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | \frac{\gamma Z}{\sigma_V} > \frac{V}{\sigma_V}\right) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | -\frac{V}{\sigma_V} > -\frac{\gamma Z}{\sigma_V}\right) \\
&= ATE - \left(\frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V}\right) \lambda\left(-\frac{\gamma Z}{\sigma_V}\right) \\
&= ATE - \left(\frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V}\right) \frac{\phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right)}{\Phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right)}
\end{aligned}
$$

$$
\begin{aligned}
TT \text{ in } P(Z) &: \\
TT\left(P(Z)\right) &\equiv E\left(Y_1 - Y_0 | D = 1, Z\right) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | \frac{V}{\sigma_V} > \frac{\gamma Z}{\sigma_V}\right) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | -\frac{V}{\sigma_V} > -\frac{\gamma Z}{\sigma_V}\right) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | -\frac{V}{\sigma_V} > \Phi^{-1}\left(1 - P(Z)\right)\right) \\
&= ATE - \left(\frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V}\right) \lambda\left(\Phi^{-1}\left(1 - P(Z)\right)\right) \\
&= ATE - \left(\frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V}\right) \frac{\phi\left(\Phi^{-1}\left(P(Z)\right)\right)}{P(Z)}
\end{aligned}
$$

Treatment parameters:

$$
\begin{aligned}
TUT \text{ in } Z &: \\
TUT(Z) &\equiv E(Y_1 - Y_0 | D = 0, Z) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | \frac{\gamma Z}{\sigma_V} < \frac{V}{\sigma_V}\right) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | \frac{V}{\sigma_V} > \frac{\gamma Z}{\sigma_V}\right) \\
&= ATE + \left(\frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V}\right) \lambda\left(\frac{\gamma Z}{\sigma_V}\right) \\
&= ATE + \left(\frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V}\right) \frac{\phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right)}{\Phi\left(-\frac{Z \cdot \gamma'}{\sigma_V}\right)}
\end{aligned}
$$

$$
\begin{aligned}
TUT \text{ in } P(Z) &: \\
TUT\left(P(Z)\right) &\equiv E\left(Y_1 - Y_0 | D = 0, Z\right) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | \frac{V}{\sigma_V} < \frac{\gamma Z}{\sigma_V}\right) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | \frac{V}{\sigma_V} > \frac{\gamma Z}{\sigma_V}\right) \\
&= ATE + \sigma_{1-0} E\left(\frac{U_1 - U_0}{\sigma_{1-0}} | \frac{V}{\sigma_V} > \Phi^{-1}\left(P(Z)\right)\right) \\
&= ATE + \left(\frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V}\right) \lambda\left(\Phi^{-1}\left(P(Z)\right)\right) \\
&= ATE + \left(\frac{\sigma_{V1} - \sigma_{V0}}{\sigma_V}\right) \frac{\phi\left(\Phi^{-1}\left(P(Z)\right)\right)}{1 - P(Z)}
\end{aligned}
$$

Matching

$$\Delta_{matching} = E(Y_1|D=1) - E(Y_0|D=0)$$

Matching (cont.)

Matching in Z:

$$
\begin{aligned}
&= ATE + E\left(U_1 | Z\gamma' > V\right) - E\left(U_0 | Z\gamma' < V\right) \\
&= ATE + E\left(U_1 | -V > -Z\gamma'\right) - E\left(U_0 | V > Z\gamma'\right) \\
&= ATE + E\left(U_1 | -\frac{V}{\sigma_V} > -\frac{Z\gamma'}{\sigma_V}\right) - E\left(U_0 | \frac{V}{\sigma_V} > \frac{Z\gamma'}{\sigma_V}\right) \\
&= ATE + \sigma_1 E\left(\frac{U_1}{\sigma_1} | -\frac{V}{\sigma_V} > -\frac{Z\gamma'}{\sigma_V}\right) - \sigma_0 E\left(\frac{U_0}{\sigma_0} | \frac{V}{\sigma_V} > \frac{Z\gamma'}{\sigma_V}\right) \\
&= ATE - \frac{\sigma_{1V}}{\sigma_V} \cdot \lambda\left(-\frac{\gamma Z}{\sigma_V}\right) - \frac{\sigma_{0V}}{\sigma_V} \cdot \lambda\left(\frac{\gamma Z}{\sigma_V}\right)
\end{aligned}
$$

Matching (cont.)

$$
\begin{aligned}
&= ATE - \frac{\sigma_{1V}}{\sigma_V} \cdot \frac{\phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right)}{\Phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right)} - \frac{\sigma_{0V}}{\sigma_V} \cdot \frac{\phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right)}{\Phi\left(-\frac{Z\cdot\gamma'}{\sigma_V}\right)} \\
&= ATE - \left(\frac{\sigma_{1V}}{\sigma_V} \cdot \frac{1}{\Phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right)} + \frac{\sigma_{0V}}{\sigma_V} \cdot \frac{1}{\Phi\left(-\frac{Z\cdot\gamma'}{\sigma_V}\right)} \right) \phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right) \\
&= ATE - \left(\frac{\frac{\sigma_{1V}}{\sigma_V} \cdot \Phi\left(-\frac{Z\cdot\gamma'}{\sigma_V}\right) + \frac{\sigma_{0V}}{\sigma_V} \cdot \Phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right)}{\Phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right)\Phi\left(-\frac{Z\cdot\gamma'}{\sigma_V}\right)} \right) \phi\left(\frac{Z\cdot\gamma'}{\sigma_V}\right)
\end{aligned}
$$

$$\Delta_{matching} = E\left(Y_1|D=1\right) - E\left(Y_0|D=0\right)$$

Matching in $P(Z)$:

$$\begin{aligned}
&= ATE + E\left(U_1|Z\gamma' > V\right) - E\left(U_0|Z\gamma' < V\right) \\
&= ATE + E\left(U_1|-V > -Z\gamma'\right) - E\left(U_0|V > Z\gamma'\right) \\
&= ATE + E\left(U_1|-\frac{V}{\sigma_V} > -\frac{Z\gamma'}{\sigma_V}\right) - E\left(U_0|\frac{V}{\sigma_V} > \frac{Z\gamma'}{\sigma_V}\right) \\
&= ATE + \sigma_1 E\left(\frac{U_1}{\sigma_1}|-\frac{V}{\sigma_V} > -\frac{Z\gamma'}{\sigma_V}\right) - \sigma_0 E\left(\frac{U_0}{\sigma_0}|\frac{V}{\sigma_V} > \frac{Z\gamma'}{\sigma_V}\right) \\
&= ATE - \frac{\sigma_{1V}}{\sigma_V}\cdot\lambda\left(-\frac{\gamma Z}{\sigma_V}\right) - \frac{\sigma_0}{\sigma_V}\cdot\lambda\left(\frac{\gamma Z}{\sigma_V}\right)
\end{aligned}$$

$$
\begin{aligned}
&= ATE - \frac{\sigma_{1V}}{\sigma_V} \cdot \lambda\left(\Phi^{-1}\left(1 - P(Z)\right)\right) - \frac{\sigma_0}{\sigma_V} \cdot \lambda\left(\Phi^{-1}\left(P(Z)\right)\right) \\
&= ATE - \frac{\sigma_{1V}}{\sigma_V} \cdot \frac{\phi\left(\Phi^{-1}\left(P(Z)\right)\right)}{P(Z)} - \frac{\sigma_0}{\sigma_V} \cdot \frac{\phi\left(\Phi^{-1}\left(P(Z)\right)\right)}{1 - P(Z)} \\
&= ATE - \left(\frac{\sigma_{1V}}{\sigma_V} \cdot \frac{1}{P(Z)} + \frac{\sigma_0}{\sigma_V} \cdot \frac{1}{1 - P(Z)}\right) \phi\left(\Phi^{-1}\left(P(Z)\right)\right) \\
&= ATE - \left(\frac{\frac{\sigma_{1V}}{\sigma_V} \cdot \left(1 - P(Z)\right) + \frac{\sigma_0}{\sigma_V} \cdot P(Z)}{P(Z)\left(1 - P(Z)\right)}\right) \phi\left(\frac{Z \cdot \gamma'}{\sigma_V}\right)
\end{aligned}
$$

Matching Bias

$$
\begin{aligned}
Bias\ \ ATE\,(Z) &= \Delta_{matching}\,(Z) - ATE\,(Z)\\
Bias\ \ MTE\,(Z) &= \Delta_{matching}\,(Z) - MTE\,(Z)\\
Bias\ \ TT\,(Z) &= \Delta_{matching}\,(Z) - TT\,(Z)\\
Bias\ \ TUT\,(Z) &= \Delta_{matching}\,(Z) - TUT\,(Z)
\end{aligned}
$$

$$
\begin{aligned}
Bias\ \ ATE\,(P(Z)) &= \Delta_{matching}\,(P\,(Z)) - ATE\,(P\,(Z))\\
Bias\ \ MTE\,(P(Z)) &= \Delta_{matching}\,(P\,(Z)) - MTE\,(P\,(Z))\\
Bias\ \ TT\,(P(Z)) &= \Delta_{matching}\,(P\,(Z)) - TT\,(P\,(Z))\\
Bias\ \ TUT\,(P(Z)) &= \Delta_{matching}\,(P\,(Z)) - TUT\,(P\,(Z))
\end{aligned}
$$

Empirical Example

$$
\begin{aligned}
Y_1 &= \mu_1 + U_1; U_1 = \alpha_{11} \cdot f_1 + \alpha_{12} \cdot f_2 + \varepsilon_1 \\
Y_0 &= \mu_0 + U_0; U_0 = \alpha_{01} \cdot f_1 + \alpha_{02} \cdot f_2 + \varepsilon_0 \\
I &= Z \cdot \gamma' - V; V = \alpha_{V1} \cdot f_1 + \alpha_{V2} \cdot f_2 + \varepsilon_V \\
D &= \mathbf{1}\left[I > 0\right]
\end{aligned}
$$

Empirical Example (cont.)

$$\left(\begin{array}{ccccc} f_1 & f_2 & \varepsilon_1 & \varepsilon_0 & \varepsilon_V \end{array} \right) \sim N\left(\mathbf{0}, \Sigma\right); \Sigma \equiv Diag\left(\begin{array}{ccccc} \sigma^2_{f_1} & \sigma^2_{f_2} & \sigma^2_V & \sigma^2_1 & \sigma^2_0 \end{array} \right)$$

$$\left[\begin{array}{c} U_1 \\ U_0 \\ V \end{array} \right] \sim N\left(\mathbf{0}, \Sigma_{U1,U0,V}\right) \equiv N\left(\mathbf{0}, \begin{array}{ccc} \sigma^2_1 & \sigma_{V1} & \sigma_{V0} \\ \cdot & \sigma^2_0 & \sigma_{10} \\ \cdot & \cdot & \sigma^2_V \end{array} \right)$$

$$\begin{array}{ll} \sigma^2_1 = \alpha^2_{11}\sigma^2_{f_1} + \alpha^2_{12}\sigma^2_{f_2} + \sigma^2_1; & \sigma_{V0} = \alpha_{V1}\alpha_{01}\sigma^2_{f_1} + \alpha_{V2}\alpha_{02}\sigma^2_{f_2} \\ \sigma^2_0 = \alpha^2_{01}\sigma^2_{f_1} + \alpha^2_{02}\sigma^2_{f_2} + \sigma^2_0; & \sigma_{10} = \alpha_{11}\alpha_{01}\sigma^2_{f_1} + \alpha_{12}\alpha_{02}\sigma^2_{f_2} \\ \sigma^2_V = \alpha^2_{V1}\sigma^2_{f_1} + \alpha^2_{V2}\sigma^2_{f_2} + \sigma^2_V; & \sigma_V = \alpha_{V1}\alpha_{11}\sigma^2_{f_1} + \alpha_{V2}\alpha_{12}\sigma^2_{f_2} \end{array}$$

Empirical Example (cont.)

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & 1 & 0 & 0 \\ \alpha_{01} & \alpha_{02} & 0 & 1 & 0 \\ \alpha_{V1} & \alpha_{V2} & 0 & 0 & 1 \end{pmatrix}$$

$$\Sigma_{U1,U0,V} \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{V1} & \sigma_{V0} \\ \cdot & \sigma_0^2 & \sigma_{10} \\ \cdot & \cdot & \sigma_V^2 \end{pmatrix} = A\Sigma A'$$

$$\begin{bmatrix} U_1 - U_0 \\ V \end{bmatrix} \sim N\left(\mathbf{0}, \begin{matrix} \sigma_{1-0}^2 & \sigma_{V1} - \sigma_{V0} \\ \cdot & \sigma_V^2 \end{matrix} \right)$$

$$\sigma_{1-0} = \sqrt{\sigma_{U1}^2 + \sigma_{U0}^2 - 2\sigma_{10}}$$

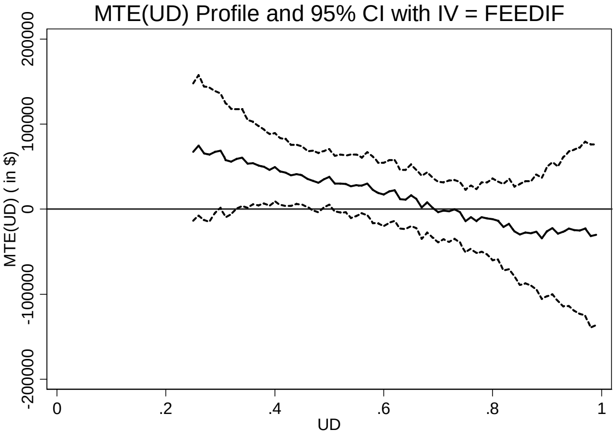
Empirical Example (cont.)

$$\mu_0 = 0; \quad \mu_0 = 1;$$
$$\alpha_{11} \text{ varies} \quad \alpha_{12} = 0.1;$$
$$\alpha_{01} = 1; \quad \alpha_{02} = 0.1;$$
$$\alpha_{V1} = 1; \quad \alpha_{V2} = 1;$$
$$\sigma_{f_1}^2 = \sigma_{f_2}^2 = \sigma_V^2 = \sigma_1^2 = \sigma_0^2 = 1$$

$$A = \begin{pmatrix} \alpha_{11} & 0.1 & 1 & 0 & 0 \\ 1 & 0.1 & 0 & 1 & 0 \\ -1 & -1 & 0 & 0 & 1 \end{pmatrix}; \Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

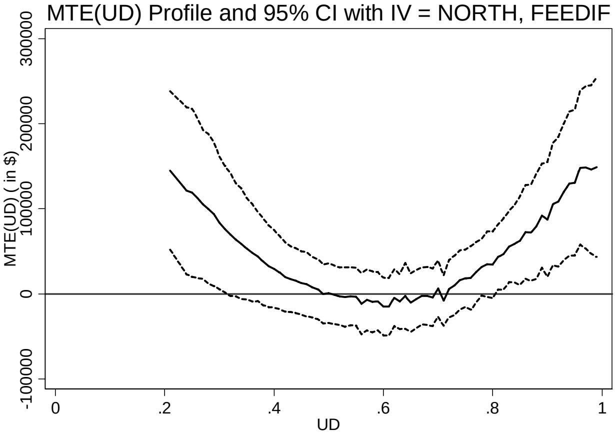
$$\Sigma_{U1,U0,V} \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{V1} & \sigma_{V0} \\ \cdot & \sigma_0^2 & \sigma_{10} \\ \cdot & \cdot & \sigma_V^2 \end{pmatrix} = A\Sigma A'$$

Example: costs of breast cancer treatments using different instruments in $P(Z)$



Source: Basu, Heckman and Urzua

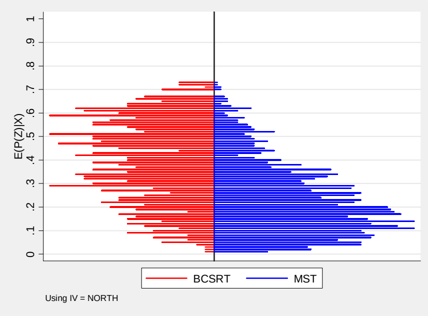
Example: costs of breast cancer treatments using different instruments in $P(Z)$



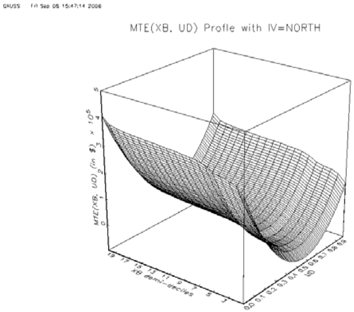
Source: Basu, Heckman and Urzua

Example: costs of breast cancer treatments using different instruments in $P(Z)$

Estimated propensity score
for BCSRT and MST



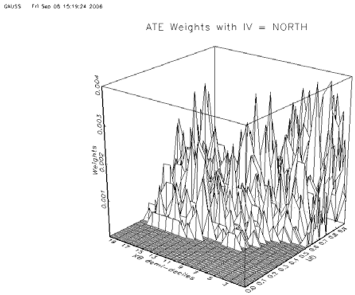
$\mathrm{MTE}(\eta_q, u_D)$



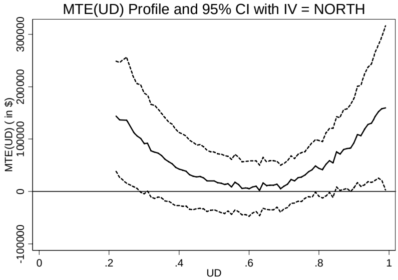
Source: Basu, Heckman and Urzua

Example: costs of breast cancer treatments using different instruments in $P(Z)$

$\omega_{\text{ATE}}(\eta_q, u_D)$



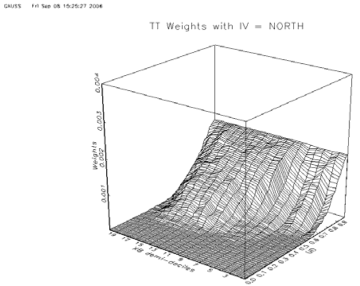
$\text{MTE}(u_D)$



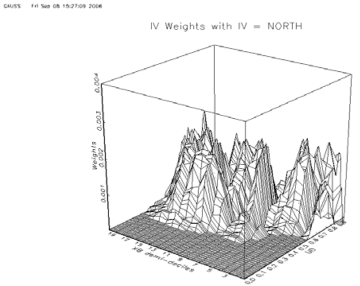
Source: Basu, Heckman and Urzua

Example: costs of breast cancer treatments using different instruments in $P(Z)$

$\omega_{\mathrm{TT}}(\eta_q, u_D)$

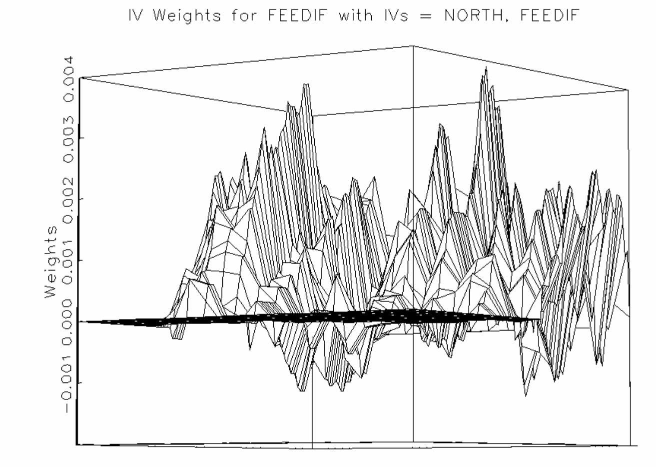


$\omega_{\mathrm{IV}}(\eta_q, u_D)$



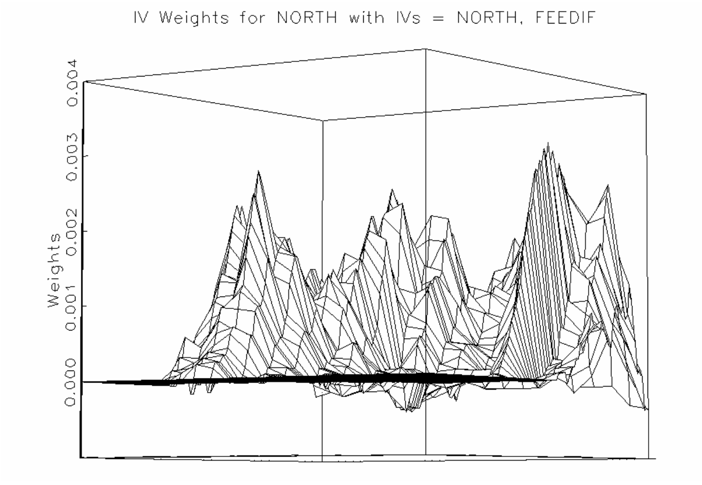
Source: Basu, Heckman and Urzua

Example: costs of breast cancer treatments using different instruments in $P(Z)$



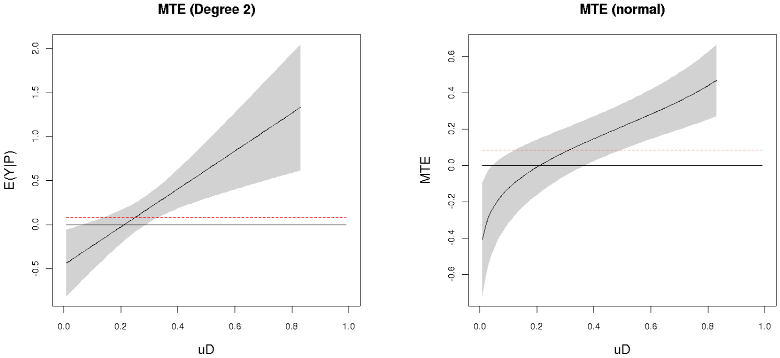
Source: Basu, Heckman and Urzua

Example: costs of breast cancer treatments using different instruments in $P(Z)$



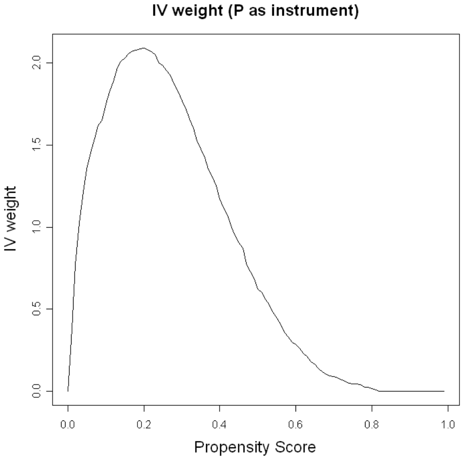
Source: Basu, Heckman and Urzua

Example: unionism on wages



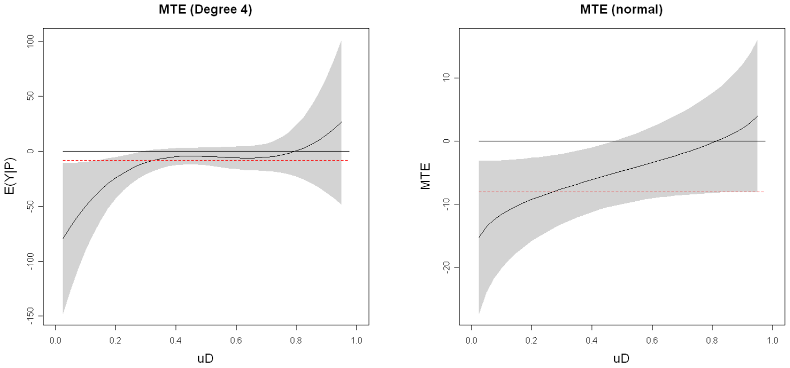
Source: Heckman, Schmierer and Urzua (2006)

Example: unionism on wages, continued



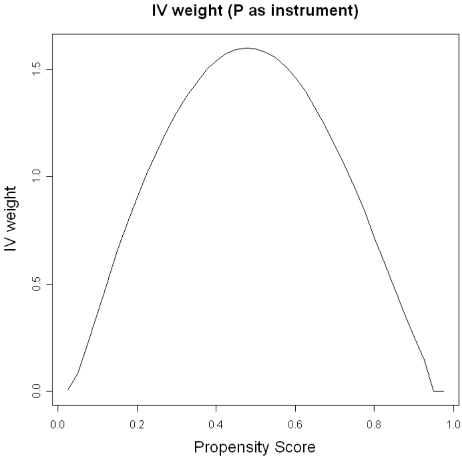
Source: Heckman, Schmierer and Urzua (2006)

Example: Chile voucher schools on test scores



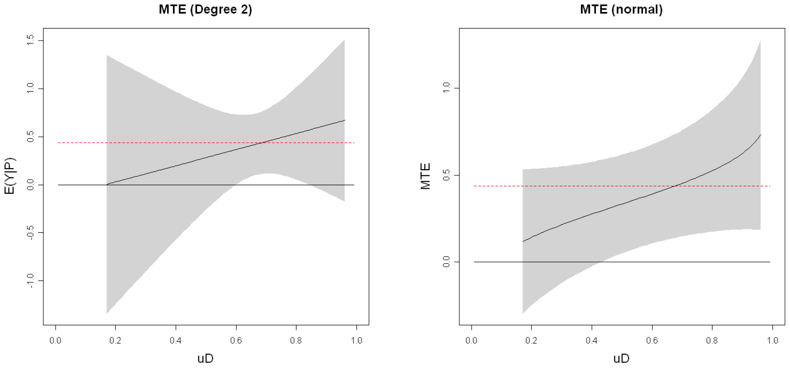
Source: Heckman, Schmierer and Urzua (2006)

Example: Chile voucher schools on test scores, continued



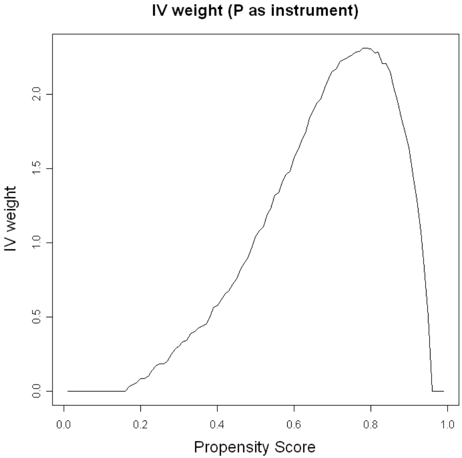
Source: Heckman, Schmierer and Urzua (2006)

Example: High school on wages



Source: Heckman, Schmierer and Urzua (2006)

Example: High school on wages, continued



Source: Heckman, Schmierer and Urzua (2006)