MARGINAL TREATMENT EFFECTS

Ivan A. Canay
Northwestern University\(^1\)
ECON 481-3


\(^1\)Today's class is based on Alex Torgovitsky's notes. I'd like to thank him for kindly sharing them.
Topics of Part I

▶ Lec I: Selection on Observables
  1. Potential Outcomes vs Latent Variables
  2. Causal Inference
  3. Selection Bias
  4. Selection on Observables & Selection on Prop. Score

▶ Lec II: Roy Models and LATE
  1. The role of heterogeneity
  2. Multiple instruments, Covariates, and Abadie’s $\kappa$

▶ Lec III: Marginal Treatment Effect
  1. Parameters as functions of MTEs
  2. Policy Relevant Treatment Effects

▶ Lec IV: Extrapolations
  1. Semi-Parametrics MTEs
  2. Weights for Target Parameters
1. Vytlacil's equivalence result
2. Marginal Treatment Effects
3. Policy Relevant Treatment Effects
AIR argue the interpretation of “error terms” is difficult to understand. “Error terms” are a way to refer to latent variables (unobservables).

They argue that assumptions about latent variables are hard to interpret.

Recall mapping from latent variable to potential outcome notation

\[ D = I\{U \leq v(Z)\} \quad \Rightarrow \quad D_z = I\{U \leq v(z)\} \]

The reverse mapping is not necessarily as easy to understand.

AIR made another argument (pg. 450) that turns out to be simply wrong: “Monotonicity has no explicit counterpart in the econometric formulation”

This statement was shown to be wrong by Vytlacil (2002).

Monotonicity is equivalent to a Roy model with a separable index.
Vytlacil’s (02) Equivalence Result

**Model 1 ("IA") as in Imbens and Angrist (1994)**

1. \( Y = DY_1 + (1 - D)Y_0 \)
2. \( D = \sum_{z \in Z} I\{Z = z\}D_z \)
3. For any \( z, z' \in Z, D_z \geq D_{z'} \) or \( D_{z'} \geq D_z \) (a.s.)
4. \( (Y_0, Y_1, \{D_z\}_{z \in Z}) \perp Z \)

**Model 2 ("ROY") is the nonparametric Roy model**

1. \( Y = DY_1 + (1 - D)Y_0 \) (same as in Model 1)
2. \( D = I\{U \leq \nu(Z)\} \) for latent variable \( U \) and unknown function \( \nu \)
3. \( (Y_0, Y_1, U) \perp Z \)

 Vytlacil (2002) showed that these two models are equivalent

 Model 2 ⇒ Model 1 is easy (next slide)

 Model 1 ⇒ Model 2 is more subtle
PROOF OF EQUIVALENCE

\[ WTS \, D = I\{U \leq v(Z)\} \quad \text{and} \quad (Y_0, Y_1, U) \perp \perp Z \quad \text{imply conditions } 1.2 - 1.4 \]
1. Vytlacil’s equivalence result
2. Marginal Treatment Effects
3. Policy Relevant Treatment Effects
Motivation

- Vytlacil’s result shows that nothing is lost from the Roy model
- Equivalence result means *same assumptions* as the “LATE framework”
- The Roy model may actually be easier to interpret for economists
- Interpret $\nu(Z) - U$ as latent utility in the choice problem for $D$
  Just keep in mind that the definition of $U$ depends on $Z$

Setup

- The analysis uses the same non-parametric Roy model as before:
  \[
  Y = DY_1 + (1 - D)Y_0 \\
  D = I\{U \leq \nu(Z)\} \quad \text{with} \quad (Y_0, Y_1, U) \perp \!\!\!\!\perp Z
  \]
- $U$ continuously distributed, *normalized to be uniform* $[0, 1]$
  Implies that $\nu(Z) = p(Z)$
- Everything is “conditional-on-$X$,” so suppress $X$ in the notation

Heckman and Vytlacil (99/05)
Normalization: Selection Equation
The Marginal Treatment Effect

Definition

- HV define the **marginal treatment effect** (MTE) as:

  \[ MTE(u) \equiv E[Y_1 - Y_0 | U = u] \]

- **MTE(u)** is the ATE for those agents with first stage unobservable \( u \)
  - Those with small \( u \) (close to 0) often choose \( D = 1 \)
  - Those with large \( u \) (close to 1) infrequently choose \( D = 1 \)

- Unobserved treatment heterogeneity if and only if non-constant MTE

Pointwise Identification of the MTE

- The MTE is point identified for all \( p \in \text{int supp } p(Z) \) (next slide):

  \[
  \frac{\partial}{\partial p} E[Y | p(Z) = p] = \text{MTE}(p)
  \]

  \[ \text{"local IV estimand"} \]

- Note that this requires \( Z \) to be continuously distributed!
Proof of identification of MTE
Proof of identification of MTE

\[
E[Y|P(Z) = p] = E[Y_1|U \leq p]p + E[Y_0|U > p](1 - p).
\]
Utility of the MTE

**AN ORGANIZING PRINCIPLE**

- $U$ provides a single dimension on which we can organize heterogeneity
- Many quantities can be written as weighted averages of the MTE

\[ \theta = \int_a^b MTE(u) \omega(u) du. \]

We will see several examples ahead

- Point identification follows from LIV if support at non-zero weights
- Tool to discuss new quantities that answer specific policy questions

**IMPLEMENTATION AND EXTRAPOLATION**

- The MTE also gives us something we can restrict and/or parameterize
- This is useful for estimation (dimension reduction)
- It is also useful for thinking about extrapolation
ATE: Unweighted Average MTE

- The ATE is the unweighted average of the MTEs:

\[
ATE = E[E[Y_1 - Y_0 | U]] = \int_0^1 MTE(u) \times 1 \, du
\]

- The ATE is point identified if \(\{0, 1\} \in \text{supp } p(Z)\)
- This follows from the LIV-MTE identification argument, e.g.:

\[
E[Y | p(Z) = 1] = E[Y_1 | U \leq 1] = E[Y_1]
\]

- Requiring \(\{0, 1\} \in \text{supp } p(Z)\) is a large support condition
- It says that there exist instrument values \(z_0, z_1 \in Z\) such that:
  - Every agent with \(z_0\) would never take the treatment (\(p(z_0) = 0\))
  - Every agent with \(z_1\) would always take the treatment (\(p(z_1) = 1\))
- A severe demand to place on the data ⇒ limited scope
- Basically means you have random assignment since \(Z = z_d \Rightarrow D = d\)
ATT/ATU: Weighted Average MTE

**ATT**

- The ATT can be written as (see problem set)

\[
ATT = \int_0^1 MTE(u) \frac{P[p(Z) \geq u]}{P[D = 1]} \, du \equiv \int_0^1 MTE(u) \omega_{ATT}(u) \, du,
\]

- Those with **low values of** \( u \) are more highly weighted
  - These are the most likely to take treatment
- The weights are known or identifiable and integrate to 1

**ATU**

- Analogous argument for the ATU:

\[
ATT = \int_0^1 MTE(u) \frac{P[p(Z) < u]}{P[D = 0]} \, du \equiv \int_0^1 MTE(u) \omega_{ATU}(u) \, du,
\]

- **High values of** \( u \) are more highly weighted (least likely to take treatment)
Suppose \( p(z) > p(z') \) for two values \( z' \) and \( z \) — then

\[
\text{LATE}_{z'} = \int_0^1 \text{MTE}(u) \omega_{\text{LATE}}(u) du, \quad \text{where} \quad \omega_{\text{LATE}}(u) \equiv \frac{I\{p(z') < u \leq p(z)\}}{p(z) - p(z')}
\]

- \( u \leq p(z') \) are always-takers for \( z' \to z \) and \( u > p(z) \) are never-takers
- \( u \in (p(z'), p(z)] \) are \( z' \to z \) compliers
- \( \text{LATE}_{z'} \) puts equal weight on compliers, 0 weight on all others
LATE as a Weighted MTE

Notice that as we take $p(z) \downarrow p(z')$,

$$\lim_{p(z) \downarrow p(z')} \text{LATE}_{z'}^z = \lim_{p(z) \downarrow p(z')} \frac{\int_{p(z')}^p \text{MTE}(u) \, du}{p(z) - p(z')} = \text{MTE}(p(z')),$$

So the MTE is a limiting (marginal) version of the LATE.
The IV Estimand as a Weighted MTE

- Suppose we use $J(Z)$ as an instrument for $D$ — IV estimand $\beta_{IV,J}$
- Using similar arguments it can be shown that

$$\beta_{IV,J} \equiv \frac{\text{cov}(J(Z), Y)}{\text{cov}(J(Z), D)} = \int_0^1 \text{MTE}(u) \omega_{IV,J}(u) du,$$

with

$$\omega_{IV,J}(u) \equiv \frac{(E[J(Z)|p(Z) \geq u] - E[J(Z)]) P[p(Z) \geq u]}{\text{cov}(J(Z), D)}.$$

- Weights are $0$ for $u < \inf \text{supp } P$ and $u > \sup \text{supp } P$
- Weights integrate to $1$
- Weights will generally be negative for some $u$:

$$E[J(Z)|p(Z) \geq u] - E[J(Z)]$$

may be both positive and negative

- Example of only positive is $J(Z) = p(Z)$ or a monotone transformation
- So IV/TLS need not estimate a “causal effect” in general
- Still consistent with the IA results since they ordered $Z$ by $p(Z)$
1. Vytlacil’s equivalence result
2. Marginal Treatment Effects
3. Policy Relevant Treatment Effects
"Policy Relevant" Parameter?

- The MTE framework partitions all agents in a clear way.
- Provides a foundation for thinking about "ideal" treatment effects.
- The "ideal" treatment effect clearly depends on the question.

- The ATE receives a lot of attention in the literature.
  But not very useful for policy — can agents still choose $D$?
- The ATT is somewhat clearer in this regard.
  Loss in benefit to treated group from discontinuing $D = 1$.

- Perhaps more relevant is changing the agent’s choice problem.
- For example, $D \in \{0, 1\}$ is attending a four-year college.
- Average effect of forcing college/no college (ATE) is not interesting.
- Nor is the effect on college-goers of shutting down college (ATT).
- More interesting are the effects via $D$ of adjusting tuition $Z$. 
HV formalize this idea as **policy relevant treatment effects** (PRTE)
- Aggregate effect on $Y$ of a change in the propensity score/instrument
- Change corresponds to a policy that affects treatment choice

Let $p^*(Z^*)$, $Z^*$ be the propensity score/instrument under a new policy
Let $D^*$ denote the treatment choice under the new policy:

$$D^* = I\{U \leq p^*(Z^*)\}$$

Letting $Y^* = D^*Y_1 + (1 - D^*)Y_0$ be the outcome under the new policy,

HV define the PRTE as:

$$\beta_{PRTE} \equiv \frac{E[Y^*] - E[Y]}{E[D^*] - E[D]}$$

The mean effect *(per net person)* of the policy change
- Implicit assumption is that the policy does not affect $(Y_0, Y_1, U)$
- Intuitively necessary — see HV for a formalization
One can show that

$$\beta_{PRTE} \equiv \frac{E[Y^*] - E[Y]}{E[D^*] - E[D]} = \int_0^1 \text{MTE}(u) \omega_{PRTE}(u) \, du$$

with

$$\omega_{PRTE}(u) \equiv \frac{F_P^-(u) - F_{P^*}^-(u)}{E[P^*] - E[P]}$$

where $F_P$ and $F_{P^*}$ are the distributions of $P \equiv p(Z)$ and $P^* \equiv p^*(Z^*)$.

$F_P^-(u) \equiv \lim_{v \uparrow u} F_P(u)$ is the left-limit of $F_P$ at $u$.

The weights show that point identifying $\beta_{PRTE}$ will be difficult.

In particular, the support of $P^*$ must be contained in that of $P$.

Since we can only possibly point identify $\text{MTE}(u)$ on the support of $P$.

Restricts to interpolating policies vs. extrapolating policies.

In addition, still need to have a continuous instrument.

Or else cannot nonparametrically point identify $\text{MTE}(u)$ for any $u$. 

The PRTE as a Weighted MTE
Instead of contrasting with status quo, could have two policies:

\[ D^a \equiv I\{U \leq p^a(Z^a)\} \quad \text{and} \quad D^b \equiv I\{U \leq p^b(Z^b)\} \]

\[ Y^a \equiv D^a Y_1 + (1 - D^a)Y_0 \quad \text{and} \quad Y^b \equiv D^b Y_1 + (1 - D^b)Y_0 \]

Then define the PRTE for \( b \) relative to \( a \) as

\[ \text{PRTE}_b^a \equiv \frac{E[Y^b] - E[Y^a]}{E[D^b] - E[D^a]} \]

Derivation of the weights just requires relabeling the previous argument

**The LATE is a PRTE**

- Policy \( a \): Every agent receives \( Z = z' \): \( p^a(\cdot) = p(\cdot), Z^a = z' \)
- Policy \( b \): Every agent receives \( Z = z \): \( p^b(\cdot) = p(\cdot), Z^b = z \)
- Then \( \text{PRTE}_b^a = \text{LATE}_{Z'}^z \)

If \( Z \) is a policy lever, the LATE may be intrinsically interesting
1. Vytlacil’s equivalence result
2. Marginal Treatment Effects
3. Policy Relevant Treatment Effects
Empirical Application

Carneiro, Heckman & Vytlacil (2011)

- Study returns to schooling in the NLSY 79 for $N = 1,747$ white males
- $Y$ is (roughly) log average hourly wages
- $D \in \{0, 1\}$ is attending some college — annualized in various ways
- $X$ contains the usual suspects plus some controls relevant for $Z$
- $Z$ are taken from a variety of other previous studies:
  1. The presence of a four-year college in the county of residence at age 14
  2. Local wage in the county of residence at age 17
  3. Local unemployment in the state of residence at age 17
  4. Average tuition in public four-year colleges in age 17 county of residence

- Analysis so far has been “conditional-on-$X$” — won’t work in practice
  Nonparametric conditioning leads to the usual curse of dimensionality
- CHV solve this by imposing some semiparametric structure
- The way they do this also helps with limited instrument support
## Variables Used in CHV (2011)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>Log wage in 1991 (average of all nonmissing wages between 1989 and 1993)</td>
</tr>
<tr>
<td>$S = 1$</td>
<td>If ever enrolled in college by 1991; zero otherwise</td>
</tr>
<tr>
<td>$Z \setminus X^b$</td>
<td>Presence of a college at age 14 (Card 1995; Stephen V. Cameron and Christopher Taber 2004), local earnings at 17 (Cameron and Heckman 1998; Cameron and Taber 2004), local unemployment at 17 (Cameron and Heckman 1998), local tuition in public four-year colleges at 17 (Thomas J. Kane and Cecilia E. Rouse 1995)</td>
</tr>
</tbody>
</table>

\(a\)We use a measure of this score corrected for the effect of schooling attained by the participant at the date of the test, since at the date the test was taken, in 1981, different individuals have different amounts of schooling and the effect of schooling on AFQT scores is important. We use a correction based on the method developed in Karsten T. Hansen, Heckman, and Kathleen J. Mullen (2004). We take the sample of white males, perform this correction, and then standardize the AFQT to have mean 0 and variance 1 within this sample. See Table A-2 in the online Appendix.

\(b\)The papers in parentheses are papers that previously used these instruments.

## Notational Differences

- Their $S$ is my $D$
- Their $Z$ is like my $(X, Z)$
- So their $Z \setminus X$ are the (excluded) instruments — what I call $Z$
- $X$ are the covariates as in my notation
Suppose that we make the following functional form assumptions:

\[ Y_d = X' \delta_d + V_d, \quad V_d = Y_d - E[Y_d|X] \quad \text{for } d = 0, 1 \]

Under these assumptions one can show that

\[ E[Y|X = x, P = p] = x' \delta_0 + px'(\delta_1 - \delta_0) + K(p, x) \]

where \( K(p, x) \equiv \int_0^p E[V_1 - V_0|X = x, U = u] \, du \)

so that \( \text{MTE}(p, x) = x'(\delta_1 - \delta_0) + \frac{\partial}{\partial p} K(p, x) \)

This is almost a **partially linear model** (Robinson 1988)

\( X \) and \( PX \) enter linearly, and through an unknown function \( K(P, X) \)

But still a **dimensionality problem**, because \( K(P, X) \) depends on \( X \)

Also, we wouldn’t be able to separately identify \( (\delta_0, \delta_1) \) from \( K(p, x) \)
To address this, CHV assume that \((X, Z) \perp \perp (V_0, V_1, U)\), so then

\[
K(p, x) \equiv \int_0^p E[V_1 - V_0 | X = x, U = u] \, du
\]

\[
= \int_0^p E[V_1 - V_0 | U = u] \, du \equiv K(p)
\]

The nonparametric component is now a function of a scalar, i.e.

\[
Y = X' \delta_0 + P X' (\delta_1 - \delta_0) + K(P) + \epsilon \quad \text{where} \quad \epsilon \equiv Y - E[Y | X, P]
\]

Now apply the classic argument for partially linear models:

\[
(*) \Rightarrow E[Y | P] = E[X | P]' \delta_0 + PE[X | P]' (\delta_1 - \delta_0) + K(P)
\]

let \( \tilde{Y} \equiv Y - E[Y | P] \) and \( \tilde{X}_j \equiv X_j - E[X_j | P] \) for each component \( j \)

then \( \tilde{Y} = \tilde{X}' \delta_0 + P \tilde{X}' (\delta_1 - \delta_0) + \epsilon \) with \( E[\epsilon | \tilde{X}, P \tilde{X}] = 0 \), linear in \( \tilde{X}, P \tilde{X} \)
Semiparametric Estimation

1. Estimate $P(X)$ using a logit or probit to get $\hat{P}(X)$
2. Construct $\tilde{Y}$ and $\tilde{X}$ through (1-dimensional!) nonparametric regression
3. Linear regression of $\tilde{Y}$ on $\tilde{X}$ and $P\tilde{X}$ to get $(\hat{\delta}_0, \hat{\delta}_1)$
4. Nonparametrically regress $Y - X'\hat{\delta}_0 - P\tilde{X}'(\hat{\delta}_1 - \hat{\delta}_0)$ on $\hat{P}$,

noting that $E[Y - X'\delta_0 - PX'(\delta_1 - \delta_0)|P] = K(P)$

Note that we want to estimate the derivative of $K$
5. Then $x'(\hat{\delta}_1 - \hat{\delta}_0) + \hat{K}'(p)$ is an estimate of $\text{MTE}(p, x)$

- $K(p, x)$ can only possibly be point identified for $(p, x) \in \text{supp}(P, X)$
- That is, for a fixed $x$, only for $p \in \text{supp}(P|X = x)$
- But $K(p)$ can be point identified for any $p \in \text{supp}(P) \supseteq \text{supp}(P|X = x)$
- So assuming $(X, Z) \perp \perp (V_0, V_1, U)$ has addressed the support issues
- As in most applications, these support issues are a big problem …
Joint Support of $P$ and Index of $X$

It is striking how small the support of $P$ is for each value of the $X$ index.

**Figure 2: Support of $P$ Conditional on $X$**

Notes: $P$ is the estimated probability of going to college. It is estimated from a logit regression of college attendance on corrected AFQT, mother’s education, number of siblings, urban residence at 14, permanent earnings in the county of residence at 17, permanent unemployment in the state of residence at 17, cohort dummies, a dummy variable indicating the presence of a college in the county of residence at age 14, average log earnings in the county of residence at age 17, and average state unemployment in the state of residence at age 17 (see Table 3). $X$ corresponds to an index of variables in the outcome equation.
Next, consider estimation of $K(P(Z))$. Equation (9) implies that

$$E(Y - X\delta_0 - P(Z)X[\delta_1 - \delta_0]|P(Z)) = K(P(Z)).$$

We thus use local polynomial regression of $Y - X\delta_0 - P(Z)X[\delta_1 - \delta_0]$ on $P(Z)$ to estimate $K(P(Z))$ and its partial derivative with respect to $P(Z)$. Local polynomial estimation not only provides a unified framework for estimating both a function and its derivative but also has a variety of desirable properties in comparison with other available nonparametric methods.\(^{19}\)

Figure 4 plots the component of the MTE that depends on $u_S$, with 90 percent confidence bands computed from the bootstrap.\(^{20}\) We fix the components of $X$ at their

\(^{19}\)Jianqing Fan and Irène Gijbels (1996) provide a detailed discussion of the properties of local polynomial estimators. In general, use of higher-order polynomials may reduce the bias but increase the variance by introducing more parameters. Fan and Gijbels suggest that the order $\pi$ of the polynomial be equal to $\pi = \tau + 1$, where $\tau$ is the order of the derivative of the function of interest that we want to fit. That is, Fan and Gijbels recommend a local linear estimator for fitting a function and a local quadratic estimator for fitting a first-order derivative. Therefore, we use a local quadratic estimator of $\partial K(P)/\partial p$. We choose the bandwidth that minimizes the residual square criterion proposed in Fan and Gijbels, which gives us a bandwidth of 0.322. Our results are robust to the choice of bandwidths between 0.1 and 0.4.

\(^{20}\)Heckman, Ichimura, and Todd (1997) show that the bootstrap provides a better approximation to the true standard errors than asymptotic standard errors for the estimation of $\beta_1$, $\beta_0$, and $K(P)$ in a model similar to the one we present here. We use 250 bootstrap replications. Throughout the paper, in each iteration of the bootstrap we re-estimate $P(Z)$ so all standard errors account for the fact that $P(Z)$ is itself an estimated object.

The support of the estimated $P(Z)$ is almost full for both $D \in \{0, 1\}$.
The assumption used in CHV is \((V_0, V_1, U) \perp \perp (X, Z)\)

- \(X \perp U\): unattractive given the variables usually included in \(X\)
- \(X\) is still allowed to have a direct effect on \(Y_d\) via \(X' \delta_d\)
- So \(X\) are completely exogenous (and correctly parameterized) covariates

### A Weaker Condition

- In fact, all that was used in the CHV derivation was:
  
  \[E[V_1 - V_0|X = x, U = u] = E[V_1 - V_0|U = u]\]

- A sufficient condition is \((V_0, V_1) \perp \perp X|U\) — does not require \(X \perp U\)

- The benefits can be seen from:

  \[
  \frac{\partial}{\partial p} \text{MTE}(p, x) = \frac{\partial}{\partial p} \left[ x' (\delta_1 - \delta_0) + \frac{\partial}{\partial p} K(p) \right] = \frac{\partial^2}{\partial p^2} K(p)
  \]

- So the **slope** of the MTE does not depend on \(x\) — **separability**
mean values in the sample. As above, we annualize the MTE. Our estimates show that, in agreement with the normal model, \( E(u_1 - u_0 | u_S = u_S) \) is declining in \( u_S \), i.e., students with high values of \( u_S \) have lower returns than those with low values of \( u_S \).

Even though the semiparametric estimate of the MTE has larger standard errors than the estimate based on the normal model, we still reject the hypothesis that its slope is zero. We have already discussed the rejection of the hypothesis that MTE is constant in \( u_S \), based on the test results reported in Table 4, panel A. But we can also directly test whether the semiparametric MTE is constant in \( u_S \) or not. We evaluate the MTE at 26 points, equally spaced between 0 and 1 (with intervals of 0.04). We construct pairs of nonoverlapping adjacent intervals (0–0.04, 0.08–0.12, 0.16–0.20, 0.24–0.28, …), and we take the mean of the MTE for each pair. These are LATEs defined over different sections of the MTE. We compare adjacent LATEs. Table 4, panel B, reports the outcome of these comparisons. For example, the first column reports that \( E(Y_1 - Y_0 | X = x, 0 \leq u_S \leq 0.04) - E(Y_1 - Y_0 | X = x, 0.08 \leq u_S \leq 0.12) = 0.0689. \)

**Figure 4.** \( E(Y_1 - Y_0 | X, U_S) \) with 90 Percent Confidence Interval—Locally Quadratic Regression Estimates

**Notes:** To estimate the function plotted here, we first use a partially linear regression of log wages on polynomials in \( X \), interactions of polynomials in \( X \) and \( P \), and \( K(P) \), a locally quadratic function of \( P \) (where \( P \) is the predicted probability of attending college), with a bandwidth of 0.32; \( X \) includes experience, current average earnings in the county of residence, current average unemployment in the state of residence, AFQT, mother’s education, number of siblings, urban residence at 14, permanent local earnings in the county of residence at 17, permanent unemployment in the state of residence at 17, and cohort dummies. The figure is generated by evaluating by the derivative of (9) at the average value of \( X \). Ninety percent standard error bands are obtained using the bootstrap (250 replications).

- This plots \( x'(\hat{\delta}_1 - \hat{\delta}_0) + \frac{\partial}{\partial p} \hat{K}(p) \) evaluated at the average of \( x \)
- Presumably only for \( p \) in its unconditional support [.032, .978]
E[s.sc/t.sc/i.sc/m.sc/a.sc/t.sc/e.sc/d MTE: N/o.sc/r.sc/m.sc/a.sc/l.sc M/o.sc/d.sc/e.sc/l.sc

2767cARnEiRO Et Al.: EStimAting mARginAl REtuRnS tO EducAtiOnVOl. 101 nO. 6

Individuals choose the schooling sector in which they have comparative advantage. The magnitude of the heterogeneity in returns on which agents select is substantial: returns can vary from $-15.6\%$ (for high US persons, who would lose from attending college) to $28.8\%$ per year of college (for low US persons).\textsuperscript{16}

The magnitude of total heterogeneity is likely to be even higher since the MTE is the average gain at that quantile of desire to attend college. In general, there will be a distribution of returns centered at each value of the MTE. Furthermore, once we account for variation in $X$ and its impact on returns through $X(\delta_1 - \delta_0)$, we observe returns as low as $-31.56\%$ and as high as $51.02\%$.

Using the weights presented in online Appendix Table A-1B, we can construct the standard treatment parameters from the MTE. We present the results in the first column of Table 5 (standard errors are bootstrapped). These include marginal returns to the three different policies considered in Table 1 (MPRTE), which are all\textsuperscript{16} One unattractive feature of the normal model is that ($\sigma_1 V$ and $\sigma_0 V$)

\[ \Delta^\text{MTE}(x, u_S) = \mu_1(x) - \mu_0(x) - (\sigma_1 V - \sigma_0 V) \Phi^{-1}(u_S), \]

where $\sigma_1 V$ and $\sigma_0 V$ are the covariances between the unobservables of the college and high school equation and the unobservable in the selection equation; and $X$ includes experience, current average earnings in the county of residence, current average unemployment in the state of residence, AFQT, mother’s education, number of siblings, urban residence at 14, permanent local earnings in the county of residence at 17, permanent unemployment in the state of residence at 17, and cohort dummies. We plot 90 percent confidence bands.

\begin{figure}[!h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{MTE Estimated from a Normal Selection Model}
\end{figure}

Notes: To estimate the function plotted here, we estimate a parametric normal selection model by maximum likelihood. The figure is computed using the following formula:

\[ \Delta^\text{MTE}(x, u_S) = \mu_1(x) - \mu_0(x) - (\sigma_1 V - \sigma_0 V) \Phi^{-1}(u_S), \]

where $\sigma_1 V$ and $\sigma_0 V$ are the covariances between the unobservables of the college and high school equation and the unobservable in the selection equation; and $X$ includes experience, current average earnings in the county of residence, AFQT, mother’s education, number of siblings, urban residence at 14, permanent local earnings in the county of residence at 17, permanent unemployment in the state of residence at 17, and cohort dummies. We plot 90 percent confidence bands.

- Notice that the normal model restricts the MTE to be monotone
- What happens as $p (u_S$ in their notation) tends to 0 or 1?
**Ho: No Unobserved Heterogeneity**

- CHV provide two tests of $H_0$: **no unobserved heterogeneity**
  
  Important null — without heterogeneity we could use simple linear IV

- Recall that under their assumptions:

  $$E[Y|X = x, P = p] = x'\delta_0 + px'(\delta_1 - \delta_0) + K(p)$$

  where

  $$K(p) \equiv \int_0^p E[V_1 - V_0|U = u] \, du$$

- No unobserved heterogeneity in treatment response if and only if

  $$E[V_1 - V_0|U = u] = C \quad \text{for all } u, \text{ some } C \iff K(p) = pC$$

- So test $H_0: E[Y|X = x, P = p]$ is linear in $p$ for each $x$

**One Way to Implement This Test**

- Assume $K(p) = \beta_0 + \beta_1p + \cdots + \beta_kp^k$ in the semiparametric procedure

- Then test $H_0: \beta_2 = \cdots = \beta_k = 0$
### Panel A. Test of linearity of $E(Y|X, P = p)$ using models with different orders of polynomials in $P$

<table>
<thead>
<tr>
<th>Degree of polynomial for model</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$-value of joint test of nonlinear terms</td>
<td>0.035</td>
<td>0.049</td>
<td>0.086</td>
<td>0.122</td>
</tr>
<tr>
<td>Adjusted critical value</td>
<td></td>
<td></td>
<td>0.057</td>
<td></td>
</tr>
<tr>
<td>Outcome of test</td>
<td></td>
<td></td>
<td>Reject</td>
<td></td>
</tr>
</tbody>
</table>

### Panel B. Test of equality of LATEs $H_0: \text{LATE}^j (U_{ij}^1, U_{ij}^2) - \text{LATE}^{j+1} (U_{ij}^{1+1}, U_{ij}^{2+1}) = 0$

<table>
<thead>
<tr>
<th>Ranges of $U_{ij}$ for LATE$^j$</th>
<th>(0.04, 0.08)</th>
<th>(0.16, 0.20)</th>
<th>(0.24, 0.28)</th>
<th>(0.32, 0.36)</th>
<th>(0.40, 0.44)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranges of $U_{ij}$ for LATE$^{j+1}$</td>
<td>(0.08, 0.12)</td>
<td>(0.16, 0.20)</td>
<td>(0.24, 0.28)</td>
<td>(0.32, 0.36)</td>
<td>(0.40, 0.44)</td>
</tr>
<tr>
<td>Difference in LATEs</td>
<td>0.0689</td>
<td>0.0629</td>
<td>0.0577</td>
<td>0.0531</td>
<td>0.0492</td>
</tr>
<tr>
<td>$p$-value</td>
<td>0.0240</td>
<td>0.0280</td>
<td>0.0280</td>
<td>0.0320</td>
<td>0.0320</td>
</tr>
<tr>
<td>Ranges of $U_{ij}$ for LATE$^j$</td>
<td>(0.48, 0.52)</td>
<td>(0.56, 0.60)</td>
<td>(0.64, 0.68)</td>
<td>(0.72, 0.76)</td>
<td>(0.80, 0.84)</td>
</tr>
<tr>
<td>Ranges of $U_{ij}$ for LATE$^{j+1}$</td>
<td>(0.56, 0.60)</td>
<td>(0.64, 0.68)</td>
<td>(0.72, 0.76)</td>
<td>(0.80, 0.84)</td>
<td>(0.88, 0.92)</td>
</tr>
<tr>
<td>Difference in LATEs</td>
<td>0.0431</td>
<td>0.0408</td>
<td>0.0385</td>
<td>0.0364</td>
<td>0.0339</td>
</tr>
<tr>
<td>$p$-value</td>
<td>0.0520</td>
<td>0.0760</td>
<td>0.0960</td>
<td>0.1320</td>
<td>0.1800</td>
</tr>
<tr>
<td>Joint $p$-value</td>
<td>0.0520</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

- **Panel B** is a more direct (maybe less clean) test of the same null hypothesis.
- **Tests the implication that MTE is constant over ranges of $[0, 1]$**
QUESTIONS?