# Marginal Treatment Effects Part II

Francis DiTraglia

Oxford Economics Summer School 2022

# Recap of Last Lecture

$$
\begin{aligned}
Y_0 &= \mu_0 + U_0 \\
Y_1 &= \mu_1 + U_1 \\
D &= 1\{\gamma_0 + \gamma_1 Z > V\} \\
Y &= (1 - D)Y_0 + DY_1
\end{aligned}
$$

$$
Z \sim \text{Bernoulli}(q) \perp\!\!\!\perp (V, U_0, U_1)
$$

$$
\begin{bmatrix} V \\ U_0 \\ U_1 \end{bmatrix} \sim \text{Normal}\left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_0\rho_0 & \sigma_1\rho_1 \\ & \sigma_0^2 & \sigma_{01} \\ & & \sigma_1^2 \end{bmatrix} \right)
$$

The Good:

- Simple model with instrument $Z \in \{0, 1\}$ and selection into treatment $D \in \{0, 1\}$.
- Treatment effects are heterogeneous and vary with "resistance" to treatment $V$.
- $\mu_0$, $\mu_1$, $\sigma_0\rho_0$, $\sigma_1\rho_1$, $q$, $\gamma_0$ and $\gamma_1$ point identified; Heckman 2-step Estimator.
- Beyond LATE: ATE, TOT, and TUT depend only on point identified parameters. . .

# Recap of Last Lecture

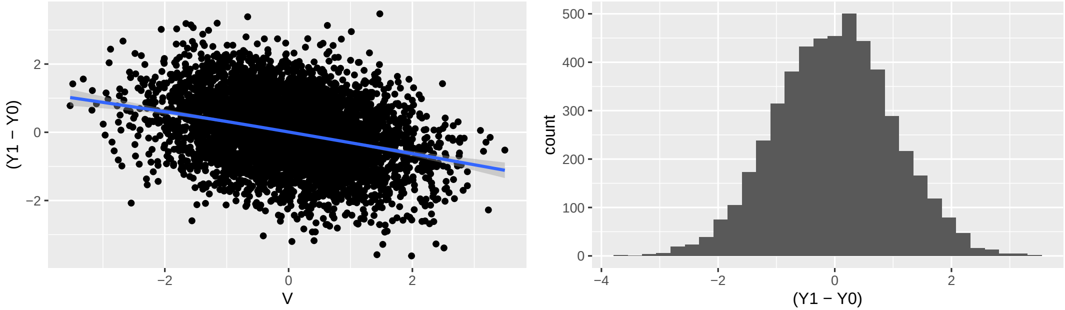
$$\text{ATE} = \mu_1 - \mu_0$$

$$\text{LATE} = \text{ATE} - (\sigma_1\rho_1 - \sigma_0\rho_0)\left[\frac{\varphi(\gamma_0+\gamma_1)-\varphi(\gamma_0)}{\Phi(\gamma_0+\gamma_1)-\Phi(\gamma_0)}\right] = \frac{\mathbb{E}(Y|Z=1)-\mathbb{E}(Y|Z=0)}{\mathbb{E}(D|Z=1)-\mathbb{E}(D|Z=0)}$$

$$\text{TOT} = \text{ATE} - (\sigma_1\rho_1 - \sigma_0\rho_0)\left[\frac{(1-q)\varphi(\gamma_0)+q\varphi(\gamma_0+\gamma_1)}{(1-q)\Phi(\gamma_0)+q\Phi(\gamma_0+\gamma_1)}\right]$$

$$\text{TUT} = \text{ATE} + (\sigma_1\rho_1 - \sigma_0\rho_0)\left[\frac{(1-q)\varphi(\gamma_0)+q\varphi(\gamma_0+\gamma_1)}{(1-q)\{1-\Phi(\gamma_0)\}+q\{1-\Phi(\gamma_0+\gamma_1)\}}\right]$$

# The Bad




## Under Normality:

1. $E(Y_1 - Y_0|V)$ is necessarily linear.
2. Unbounded ATEs for people with "extreme" values of $V$.

# Relaxing Normality: the Latent Index Selection Model (LISM)

$$
\begin{aligned}
Y_0 &= \mu_0 + U_0 & \qquad Y &= (1-D)Y_0 + DY_1 \\
Y_1 &= \mu_1 + U_1 & \qquad Z &\sim \text{Bernoulli}(q) \perp\!\!\!\perp (V, U_0, U_1) \\
D &= 1\{\gamma_0 + \gamma_1 Z > V\} & \qquad \mathbb{E}(V) &= \mathbb{E}(U_0) = \mathbb{E}(U_1) = 0
\end{aligned}
$$

## The Good:

- Simple model with instrument $Z \in \{0,1\}$ and selection into treatment $D \in \{0,1\}$.
- Treatment effects are heterogeneous and vary with "resistance" to treatment $V$.
- No longer assume that $(U_0, U_1, V)$ are jointly normal; mean zero WLOG.

## Questions

1. How does this compare to the LATE model?
2. Is this model identified? If so can we estimate it?
3. If we can estimate it, does it allow us to go beyond late to ATE, TUT, TOT etc?

# Assumptions of the Latent Index Selection Model

### Treatment Take-up

$D(Z) = 1\{\gamma_0 + \gamma_1 Z > V\}$

### Instrument Relevance

$\mathbb{P}(\gamma_0 > V) \neq \mathbb{P}(\gamma_0 + \gamma_1 > V)$

### Instrument Exogeneity

$Z \perp\!\!\!\perp (V, Y_0, Y_1)$

- $\gamma_0 + \gamma_1 Z$ is called the "latent index"
- We used relevance implicitly in our Heckman Two-step procedure.
- $Z \perp\!\!\!\perp (V, U_0, U_1)$, $Y_0 = \mu_0 + U_0$, $Y_1 = \mu_1 + U_1 \implies$ exogeneity

# Potential Treatments

- We described LATE model using "compliance type" variable $T \in \{n, a, c, d\}$
- Equivalently, can describe using "potential treatments," a binary encoding: $(D_0, D_1)$

$$
\begin{array}{rcccccc}
\text{Never-taker:} & T = n & \iff & D(Z) = 0 & \iff & (D_0 = 0, D_1 = 0) \\
\text{Always-taker:} & T = a & \iff & D(Z) = 1 & \iff & (D_0 = 1, D_1 = 1) \\
\text{Complier:} & T = c & \iff & D(Z) = Z & \iff & (D_0 = 0, D_1 = 1) \\
\text{Defier:} & T = d & \iff & D(Z) = (1 - Z) & \iff & (D_0 = 1, D_1 = 0)
\end{array}
$$

## No Defiers aka Monotonicity

$\mathbb{P}(T = d) = 0 \iff$ either $D_0 \leq D_1$ or $D_1 \leq D_0$ with probability one.

## Unconfounded Type

$Z \perp\!\!\!\perp T \iff Z \perp\!\!\!\perp (D_0, D_1)$

# (Slightly) Stronger Version of LATE Assumptions

### Existence of Compliers in terms of Observables

$\mathbb{P}(T = c) > 0 \iff \mathbb{E}[D|Z = 1] \neq \mathbb{E}[D|Z = 0]$

### No Defiers in terms of Potential Treatments

Either $D_0 \leq D_1$ or $D_1 \leq D_0$ with probability one.

### Replacement for Mean Exclusion

$Z \perp\!\!\!\perp (Y_0, Y_1, D_0, D_1)$

- Equivalent to $Z \perp\!\!\!\perp (Y_0, Y_1, T)$
- Implies $Z \perp\!\!\!\perp (D_0, D_1)$, which is equivalent to unconfounded type.
- Implies but is slightly stronger than mean exclusion.

# These two models are equivalent!

**Latent Index Selection Model**

1. $D = 1\{\gamma_0 + \gamma_1 Z > V\}$
2. $\mathbb{P}(\gamma_0 > V) \neq \mathbb{P}(\gamma_0 + \gamma_1 > V)$
3. $Z \perp\!\!\!\perp (Y_0, Y_1, V)$

**Local Average Treatment Effects Model**

1. Either $D_0 \leq D_1$ or $D_1 \leq D_0$ wp 1.
2. $\mathbb{E}[D|Z = 1] \neq \mathbb{E}[D|Z = 0]$
3. $Z \perp\!\!\!\perp (Y_0, Y_1, D_0, D_1)$

**LISM Assumptions $\Rightarrow$ LATE Assumptions**

- Straightforward. Details follow on the next slide.

**LATE Assumptions $\Rightarrow$ LISM Assumptions**

- A bit trickier. See: Glickman & Normand (2000) and Vytacil (2002)

## $Z \perp\!\!\!\perp (Y_0, Y_1, D_0, D_1)$

- $D = 1\{\gamma_0 + \gamma_1 Z > V\} \implies (D_0, D_1)$ are a function of $V$.
- In particular: $D_0 \equiv D(Z=0) = 1\{\gamma_0 > V\}$, $D_1 \equiv D(Z=1) = 1\{\gamma_0 + \gamma_1 > V\}$
- The LISM assumes $Z \perp\!\!\!\perp (Y_0, Y_1, V)$, so by Decomposition: $Z \perp\!\!\!\perp (Y_0, Y_1, D_0, D_1)$.

## $\mathbb{P}(D=1|Z=1) \neq \mathbb{P}(D=1|Z=0)$

- The LISM assumes that $\mathbb{P}(\gamma_0 > V) \neq \mathbb{P}(\gamma_0 + \gamma_1 > V)$
- $\mathbb{P}(D=1|Z=0) = \mathbb{P}(\gamma_0 > V)$, $\mathbb{P}(D=1|Z=1) = \mathbb{P}(\gamma_0 + \gamma_1 > V)$

## Either $D_0 \leq D_1$ or $D_1 \leq D_0$ with probability one.

- $\mathbb{P}(\gamma_0 > V) \neq \mathbb{P}(\gamma_0 + \gamma_1 > V)$ rules out $\gamma_1 = 0$.
- $\gamma_1 > 0 \Rightarrow \gamma_0 + \gamma_1 > \gamma_0 \Rightarrow \mathbb{P}(D_0 \leq D_1) = \mathbb{P}(1\{\gamma_0 > V\} \leq 1\{\gamma_0 + \gamma_1 > V\}) = 1$
- $\gamma_1 < 0 \Rightarrow \gamma_0 + \gamma_1 < \gamma_0 \Rightarrow \mathbb{P}(D_1 \leq D_0) = \mathbb{P}(1\{\gamma_0 + \gamma_1 > V\} \leq 1\{\gamma_0 > V\}) = 1$

# The Generalized Roy Model

## Model

$$Y_0 = \mu_0(X) + U_0$$
$$Y_1 = \mu_1(X) + U_1$$
$$Y = (1 - D)Y_0 + DY_1$$

## Assumptions

1. $D = 1\{\nu(X, Z) > V\}$
2. $Z \perp\!\!\!\perp (Y_0, Y_1, V)|X$
3. Distribution of $V|X = x$ is continuous.

- Covariates $X$: **observed heterogeneity**; $(U_0, U_1, V)$: **unobserved heterogeneity**
- $U_0 \equiv Y_0 - \mathbb{E}(Y_0|X)$; $U_1 \equiv Y_1 - \mathbb{E}(Y_1|X)$ so both are mean zero.
- $Z$ may not be be binary; unknown function $\nu(\cdot)$

# Monotonicity

## Model

$$Y_0 = \mu_0(X) + U_0$$
$$Y_1 = \mu_1(X) + U_1$$
$$Y = (1 - D)Y_0 + DY_1$$

## Assumptions

1. $D = 1\{\nu(X, Z) > V\}$
2. $Z \perp\!\!\!\perp (Y_0, Y_1, V)|X$
3. Distribution of $V|X = x$ is continuous.

- Holding $X$ fixed, we can shift $\nu(X, Z)$ by changing $Z$ *without affecting* $V$.
- Why? Conditional on $X$, $Z$ and $V$ are independent and $V$ doesn't enter $\nu(\cdot)$.
- For a given shift in $Z$, two people with the same observed characteristics $X$ experience the same shift in $\nu(\cdot)$ **regardless of whether they have different resistance to treatment** $V$

# Normalization: Transform $V$ to Uniform$(0,1)$

- For any continuous RV $W$ with CDF $H$, $\widetilde{W} \equiv H(W) \sim \text{Uniform}(0,1)$
- Condition on $(X = x)$; let $F_x$ be the conditional dist of $V|X = x$ (continuous)
- Remember: conditional on $X$, $Z$ and $V$ are independent!

$$\begin{aligned} D|(X = x) &= 1\left\{\nu(x, Z) > V\right\} = 1\{F_x\left(\nu(x, Z)\right) > F_x(V)\} \\ &= 1\{F_x(\nu(x, Z)) > \widetilde{V}\} = 1\{g(x, Z) > \text{Uniform}\} \end{aligned}$$

- If $W \sim \text{Uniform}(0,1)$ then $\mathbb{P}(W < c) = c$.

$$\pi(x, z) \equiv \mathbb{P}(D = 1|X = x, Z = z) = \mathbb{P}(g(x, z) > \text{Uniform}) = g(x, z)$$

- **WLOG normalize** $V|X = x \sim \text{Uniform}(0,1) \implies V|(X = x, Z = z)$ also uniform
- The function $\nu(\cdot)$ becomes the **propensity score** $\pi(X, Z)$.

# Generalized Roy Model

## Model

$$
\begin{aligned}
Y_0 &= \mu_0(X) + U_0 \\
Y_1 &= \mu_1(X) + U_1 \\
Y &= (1 - D)Y_0 + DY_1 \\
\pi(X, Z) &\equiv \mathbb{P}(D = 1|X, Z)
\end{aligned}
$$

## Assumptions

1. $D = 1\{\pi(X, Z) > V\}$
2. $Z \perp\!\!\!\perp (Y_0, Y_1, V)|X$
3. $V|(X = x, Z = z) \sim \text{Uniform}(0, 1)$

# ATE, TOT and TUT in the Generalized Roy Model

$$
\begin{aligned}
\text{ATE}(x) &\equiv \mathbb{E}[Y_1 - Y_0 | X = x] = \mu_1(x) - \mu_0(x) \\
\text{TOT}(x) &\equiv \mathbb{E}[Y_1 - Y_0 | X = x, D = 1] = \mu_1(x) - \mu_0(x) + \mathbb{E}[U_1 - U_0 | X = x, D = 1] \\
\text{TUT}(x) &\equiv \mathbb{E}[Y_1 - Y_0 | X = x, D = 0] = \mu_1(x) - \mu_0(x) + \mathbb{E}[U_1 - U_0 | X = x, D = 0]
\end{aligned}
$$

- Same definitions as before, but now we are conditioning on $X$.
- Average over the distribution of $X$ to obtain unconditional versions.

# Policy-Relevant Treatment Effects (PRTEs)

$$\text{PRTE}(x) \equiv \frac{\mathbb{E}[Y_i|X_i = x, \text{New Policy}] - \mathbb{E}[Y_i|X_i = x, \text{Old Policy}]}{\mathbb{E}[D_i|X_i = x, \text{New Policy}] - \mathbb{E}[D_i|X_i = x, \text{Old Policy}]}$$

- Compare a new policy to old one; average over $X$ to obtain unconditional version.
- Policy $\equiv$ change in the propensity score $\pi(Z, X)$ that changes who is treated without affecting $(Y_1, Y_0, V)$.
- PRTE is the average change in $Y$ *per person shifted into treatment*.
- At some values of $x$, people may be shifted *out of treatment*
- A LATE is a PRTE, but a given LATE may not answer *your* policy question!

# Marginal Treatment Effects (MTEs)

## Textbook Normal Model

- Any treatment effect of interest can be calculated from $(\gamma_0, \gamma_1, \mu_0, \mu_1, \delta)$.
- These parameters are identified: Heckman Two-step approach

## Generalized Roy Model

- Any treatment effect can be calculated as from knowledge of the **Marginal Treatment Effect** (MTE) function

$$\text{MTE}(v, x) \equiv \mathbb{E}(Y_1 - Y_0 | X = x, V = v)$$

- How do treatment effects vary with observed $(x)$ and unobserved $(v)$ heterogeneity?
- No unobserved heterogeneity $\implies$ MTE is *constant* as a function of $v$.
- Like textbook model parameters, MTE does *not* depend on the instrument $Z$.

# From MTE Function to Target Parameters

## Target Parameters

- ATE, TOT, TUT, PRTEs, LATE, etc.

## General Approach

- Any of the above (and more!) can be computed as a weighted average of the MTE.

## Example: ATE from MTE

$$\begin{aligned}\text{ATE}(x) &\equiv \mathbb{E}[Y_1 - Y_0 | X = x] = \mathbb{E}_{V|X=x}[\mathbb{E}(Y_1 - Y_0 | X = x, V = v)] \\ &= \mathbb{E}_{V|X=x}[\text{MTE}(X, V)] = \int \text{MTE}(x, v)\, dF_{V|X=x}(v) \\ &= \int_0^1 \text{MTE}(v, x) \times 1\, dv\end{aligned}$$

- Follows because $V|X = x \sim \text{Uniform}(0, 1)$.
- See Mogstad & Torgovitsky (2018) for other weighting functions.

# How can we identify the MTE function?
## Notation

$$m(p,x) \equiv \mathbb{E}\left[Y|\pi(X,Z)=p, X=x\right]$$
$$m_0(p,x) \equiv \mathbb{E}\left[Y|\pi(X,Z)=p, X=x, D=0\right]$$
$$m_1(p,x) \equiv \mathbb{E}\left[Y|\pi(X,Z)=p, X=x, D=1\right]$$

## Two Approaches

1. Local Instrumental Variables

$$\text{MTE}(p,x) = \frac{\partial}{\partial p} m(p,x)$$

2. Separate Estimation

$$\text{MTE}(p,x) = [m_0(p,x) - m_1(p,x)] + p\frac{\partial}{\partial p} m_1(p,x) + (1-p)\frac{\partial}{\partial p} m_0(p,x)$$

# The Local Instrumental Variables Approach

## Can Show that

$$m(p, x) \equiv E[Y|\pi(X, Z) = p, X = x] = \mu_0(x) + p[\mu_1(x) - \mu_0(x)] + K(p, x)$$
$$K(p, x) \equiv pE(U_1 - U_0|V \leq p, X = x) = \int_0^p E(U_1 - U_0|X = x, V = v)\, dv.$$

## Differentiating with respect to $p$

$$\begin{aligned}
\frac{\partial}{\partial p} E[Y|P(X, Z) = p, X = x] &= \mu_1(x) - \mu_0(x) + \frac{\partial}{\partial p} K(p, x) \\
&= \mu_1(x) - \mu_0(x) + E(U_1 - U_0|X = x, V = p) \\
&\equiv \mathrm{MTE}(p, x)
\end{aligned}$$

- 2nd-to-last equality: definition of $K(p, x)$ and Fundamental Theorem of Calculus.

# Theory Versus Practice

- Both local IV and separate estimation approaches involve non-parametric regression of $Y$ on $X$ and the propensity score.
- This is extremely challenging in practice even if $X$ is low-dimensional!
- Need variation in propensity score for fixed $X$; this comes from $Z$.
- To non-parametrically identify the full MTE function, need an instrument that allows $\pi(X, Z)$ to vary over the **full range** $[0, 1]$ for any value of $X$!
- In practice, researchers make simplifying assumptions and carry out semi-parametric or flexible parametric estimation.
- This **invariably involves interpolation / extrapolation** to some degree!
- See Mogstad & Torgovitsky (2018) for a partial identification approach.

# Cornelissen et al (QJE; 2018) - Who Benefits from Universal Child Care?

## Background

- Major policy question: causal effect of early childhood interventions, including state-provided day care.
- Some studies of highly-targeted programs (e.g. Head Start / Perry Preschool) find sizable positive effects.
- Evidence for *universal* provision is mixed: some find sizable *negative* effects (Quebec study).
- How to rationalize these conflicting findings?
- Maybe targeted programs enroll children *most likely to benefit*, i.e. those with an adverse home environment.

# Cornelissen et al (QJE; 2018) - Who Benefits from Universal Child Care?

## This Study

- Study provision of universal preschool/childcare in Germany using MTE approach.
- Treatment is **early attendance**, defined as attending for *at least* three years.
- Instrument is a staggered roll-out of 1990s policy reform that affected the number of slots for publicly-provided childcare in different places.
- Main outcome is a universal school readiness exam administered at age 6.

# Cornelissen et al (QJE; 2018) - Who Benefits from Universal Child Care?

## Main Findings

- Evidence of *reverse selection on gains* from observed characteristics.
- Minorities benefit most from childcare but are least likely to enroll.
- Similar selection on unobservables: "high resistance" children benefit most.
- Effect is so strong that TUT $>$ ATE $> 0 >$ TOT!
- Evidence that treatment effect heterogeneity comes from $Y_0$ rather than $Y_1$.

## The Rest of the Lecture

- We'll focus on their **implementation** of MTE methods.
- Also talk a bit about policy counterfactuals.
- See the paper for more details.

# A Simplified MTE Model

## Additive Separability

- $\mathbb{E}[U_0|V, X] = \mathbb{E}[U_0|V]$ and $\mathbb{E}[U_1|V, X] = \mathbb{E}[U_1|V]$
- Changing $X$ only affects the *intercept* of the MTE, viewed as a function of $v$.
- Still allows $V$ to vary with $X$.

## Linearity

- $\mathbb{E}[Y_0|X = x] = x'\beta_0$ and $\mathbb{E}[Y_1|X = x] = x'\beta_1$
- Restricts the way that covariates affect the intercept of the MTE function.

# Implications of Separability and Linearity

## MTE Function

$$\begin{aligned}
\text{MTE}(p,x) &= \mu_1(x) - \mu_0(x) + E(U_1 - U_0 | X = x, V = p) \\
&= \mu_1(x) - \mu_0(x) + E(U_1 - U_0 | V = p) && \text{(Separability)} \\
&= x'(\beta_1 - \beta_0) + E(U_1 - U_0 | V = p) && \text{(Linearity)} \\
&= x'(\beta_1 - \beta_0) + \frac{d}{dp} K(p) && \text{(Linearity)}
\end{aligned}$$

## Observed Conditional Mean Function

$$\begin{aligned}
\mathbb{E}[Y|\pi(X,Z) = p, X = x] &= \mu_0(x) + p[\mu_1(x) - \mu_0(x)] + K(p,x) \\
&= x'\beta_0 + x'(\beta_1 - \beta_0)p + K(p)
\end{aligned}$$

- This is a **semi-parametric model**: linear regression plus unknown function $K(p)$

# A Parametric Approximation

- Could choose to carry out semi-parametric estimation, but Cornelissen et al (2018) take a simpler approach.
- Model $K(p)$ as a polynomial in $p$; don't include constant or first-order term since they're already in the regression:

$$\mathbb{E}[Y|\pi(X,Z) = p, X = x] = x'\beta_0 + x'(\beta_1 - \beta_0)p + \sum_{j=2}^{J} \alpha_j p^j$$

- If we knew $p$, we could run this regression; unfortunately we don't know it!

# Implementation

1. Run probit/logit of $D_i$ on $(X_i, Z_i)$ to estimate the propensity scores $\widehat{p}_i$.
2. Estimate $\beta_0, \beta_1, \alpha$ from the following regression:

$$Y_i = X_i\beta_0 + X_i'(\beta_1 - \beta_0)\widehat{p}_i + \sum_{j=2}^{J} \alpha_j \widehat{p}_i^j + \epsilon_i$$

3. Construct the estimated MTE function as follows:

$$\widehat{\text{MTE}}(p, x) = \frac{\partial}{\partial p}\left[x'\widehat{\beta}_0 + x'(\widehat{\beta}_1 - \widehat{\beta}_0)p + \sum_{j=2}^{J} \widehat{\alpha}_j p^j\right]$$

4. Take weighted average of $\widehat{\text{MTE}}(p, x)$ to construct desired target parameter.

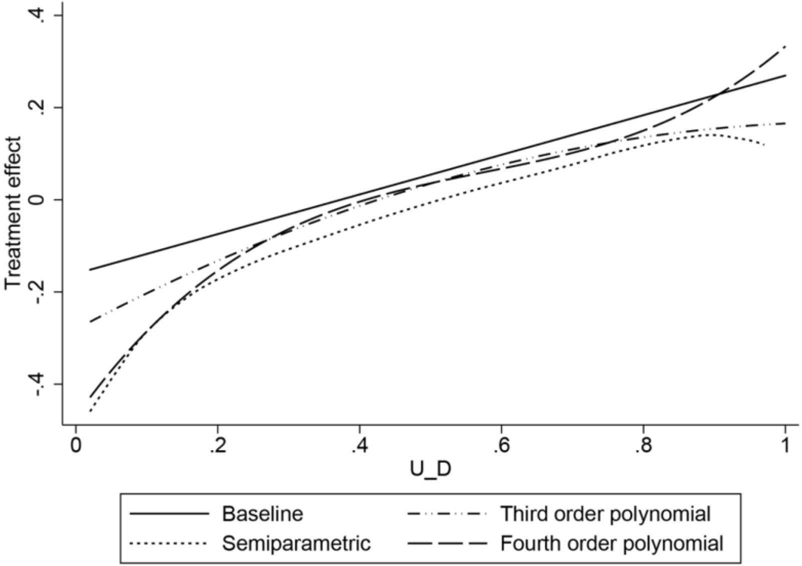
# Some Specifics from Cornelissen et al (2018)

- Add municipality ($R$) and exam cohort ($T$) dummies:

$$Y = X\beta_0 + \alpha R + \tau T + X(\beta_1 - \beta_0)\widehat{p} + \sum_{j=1}^{J} \alpha_j \widehat{p}^j + \epsilon$$

- Experiment with $J = 2$, $J = 3$, $J = 4$, and a semi-parametric specification.
- Remember: we *differentiate* to get the MTE, so $J = 2$ is a *linear* specification for $\mathbb{E}(U_1 - U_0|V)$. Sound familiar?
- Similar results across the different specifications of $K(p)$ in this case.

# Treatment effects **increase** with resistance to treatment!



# Policy Counterfactuals

TABLE 9
POLICY-RELEVANT TREATMENT EFFECTS

| | PRTE (1) | Propensity Score: Baseline (2) | Propensity Score: Policy (3) |
|---|---|---|---|
| 1. Bring 2002 $P(Z)$ to .9 by adding .275 | .160* (.085) | .67 | .90 |
| 2. Bring 2002 $P(Z)$ to .9 by multiplying 1.5 | .165* (.087) | .67 | .90 |
| 3. Lift 2002 cohort's coverage rate ($Z$) to 1 if < 1 | .123 (.077) | .67 | .71 |
| 4. Add .4 to 2002 cohort's coverage rate ($Z$) | .141* (.086) | .67 | .72 |