Treatment Effects Part 1

Richard L. Sweeney

based on slides by Chris Conlon

Empirical Methods
Spring 2021

Overview

This lecture draws heavily upon

- 2012 AEA continuing education lectures by Imbens and Woodldridge (full materials available here.)
- Mostly Harmless Econometrics
- The Mixtape (read online here)

The Evaluation Problem

- The issue we are concerned about is identifying the effect of a policy or an investment or some individual action on one or more outcomes of interest
- This has become the workhorse approach of the applied microeconomics fields (Public, Labor, etc.)
- Examples may include:
  - The effect of taxes on labor supply
  - The effect of education on wages
  - The effect of incarceration on recidivism
  - The effect of competition between schools on schooling quality
  - The effect of price cap regulation on consumer welfare
  - The effect of indirect taxes on demand
  - The effects of environmental regulation on incomes
  - The effects of labor market regulation and minimum wages on wages and employment

Setup

Typically attributed to Rubin

- Observe N units, indexed by i, drawn randomly from a larger population
- Postulate two potential outcomes for each unit $\{Y_i(1), Y_i(0)\}$ depending on whether they receive treatment or not.
- Observe additional *exogenous* covariates X_i
- Consider a binary treatment W_i such that

$$Y_i \equiv Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0 \\ Y_i(1) & \text{if } W_i = 1 \end{cases}$$

SUTVA

- Note there is already an important assumption embedded in this setup, the stable unit treatment value assumption (SUTVA).
- Assume that the outcome, in either state for unit i does not depend on the assignment of other units.
- This is likely to fail in many important settings. Examples?

Methods

1. Matching
2. Instrumental Variables
3. Difference in Difference and Natural Experiments
4. RCTs
5. Structural Models

- Key distinction: the treatment effect of some program (a number) from understanding how and why things work (the mechanism).
- Models let us link numbers to mechanisms.

The Evaluation Problem

Two major problems:

- All individuals have different treatment effects τ_i (heterogeneity).
- Individual treatment effects $\tau_i = Y_{1i} - Y_{0i}$ are never observed (FPOCI)

The Evaluation Problem

Two major problems:

- All individuals have different treatment effects τ_i (heterogeneity).
- Individual treatment effects $\tau_i = Y_{1i} - Y_{0i}$ are never observed (FPOCI)

What is hard here?

- Selection in treatment may be endogenous. That is W_i depends on $Y_i(1), Y_i(0)$.
- Fisher or Roy (1951) model:

$$Y_i = (Y_i(1) - Y_i(0))W_i + Y_i(0) = \alpha + \tau_i W_i + u_i$$

- Agents usually choose W_i with τ_i or u_i in mind.
- Can't necessarily pool across individuals since τ_i is not constant.

Structural vs. Reduced Form

- Usually we are interested in one or two parameters of the distribution of τ_i (such as the average treatment effect or average treatment on the treated).
- Most program evaluation approaches seek to identify one effect or the other effect in reduced form, using quasi-experimental variation.
- The structural approach attempts to recover the entire joint $f(\tau, u_i)$ distribution but generally requires more assumptions, but then we can calculate whatever we need.
- Instead we often focus on simpler estimands.

Common Objects of Interest

- Population average treatment effect (PATE)

$$\tau_P = E\left[Y_i(1) - Y_i(0)\right]$$

- Population average treatment effect for treated units (PATT)

$$\tau_{P,T} = E\left[Y_i(1) - Y_i(0)|W = 1\right]$$

- Sample average treatment effect (SATE)

$$\tau_S = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$$

- Sample average treatment effect for treated units (SATT)

$$\tau_{S,T} = \frac{1}{N_T}\sum_{i \in W_i=1}(Y_i(1) - Y_i(0))$$

Estimation under unconfoundedness

Assumption: 1

$$(Y_i(0), Y_i(1)) \perp W_i | X_i$$

- Sometimes called "conditional independence assumption" or "selection on observables".
- Can see this is implicit in the regression $Y_i = \alpha + \tau W_i + X_i'\beta + \epsilon_i$ where $\epsilon_i \perp X_i$ under the assumption of a constant treatment effect (otherwise this is not the same)

Assumption 2 (Overlap)

$$0 < Pr(W_i = 1 | X_i) < 1$$

How useful are these assumptions?

Imbens (2015) has a good discussion on this. Suggests following motivations:

- This is a natural starting point. Compare treatment and control units, after adjusting for observables. Need not be the last word!
- *All* comparisons involve comparing treated to untreated units. Absent RCT, its up to researcher to investigate which comparisons to emphasize
- Often specifying a model can clarify how sensible this is. Guido has a good example on costs in the paper.

Under these assumptions, can we just use regression?

- Let $\mu_w(x) = E[Y_i(w)|X_i = x]$
- A regression estimate of τ is then

$$\hat{\tau}_{reg} = \frac{1}{N}\sum_i W_i(Y_i - \hat{\mu}_0(X_i)) + (1 - W_i)(\hat{\mu}_1(X_i) - Y_i)$$

- Typically estimate

$$Y_i = \alpha + \beta' X_i + \tau W_i + \epsilon_i$$

  which assumes $\mu_w(x) = \beta' x + \tau * w$
- Could easily also compute

$$\mu_w(x) = \alpha_w + \beta'_w x$$

- Key point is that this estimator can be viewed as a missing data problem, where predictions are computed using regression.

When is this likely to be a problem?

- Note $\mu_0(x)$ is used to predict the "missing" control outcomes for the treated observations.
- Want this predition at the average treated covariates $\bar{X}_T$
- With linear regression, our average control prediction for the treated observations is going to be $\bar{Y}_C + \hat{\beta}'(\bar{X}_T - \bar{X}_C)$
- Ok if:
  1. $\mu()$ is properly specified
  2. treated and control observations are similar (in X)
- First condition is untestable, but in practice predictions are often sensitive to functional form
- Leads to a big emphasis on covariate balance.

Matching

- Regression imputes missing potential outcomes using regression.
- Matching imputes using the *realized* outcome of (nearly) identical units in the opposite assignment group.
- Remember, we're in a world where we've assumed unconfoundedness. Only challenge is that the treatment group and the control group don't have the same distribution of X's.
- Re-weight the un-treated population so that it resembles the treated population.
- Once distribution of X_i is the same for both groups $X_i|W_i \sim X_i$ then we assume all other differences are irrelevant and can just compare means.

Matching

Let $F^1(x)$ be the distribution of characteristics in the treatment group, we can define the ATE as

$$\begin{aligned} & E[Y(1) - Y(0)|T = 1] \\ = \; & E_{F^1(x)}[E(Y(1) - Y(0)|T = 1, X)] \\ = \; & E_{F^1(x)}[E(Y(1)|T = 1, X)] - E_{F^1(x)}[E(Y(0)|T = 1, X)] \end{aligned}$$

The first part we observe directly:

$$= \; E_{F^1(x)}[E(Y(1)|T = 1, X)]$$

But the counterfactual mean is not observed!

$$= \; E_{F^1(x)}[E(Y(0)|T = 1, X)]$$

But conditional independence does this for us:

$$E_{F^1(x)}[E(Y(0)|T = 1, X)] = E_{F^1(x)}[E(Y(0)|T = 0, X)]$$

A Matching Example

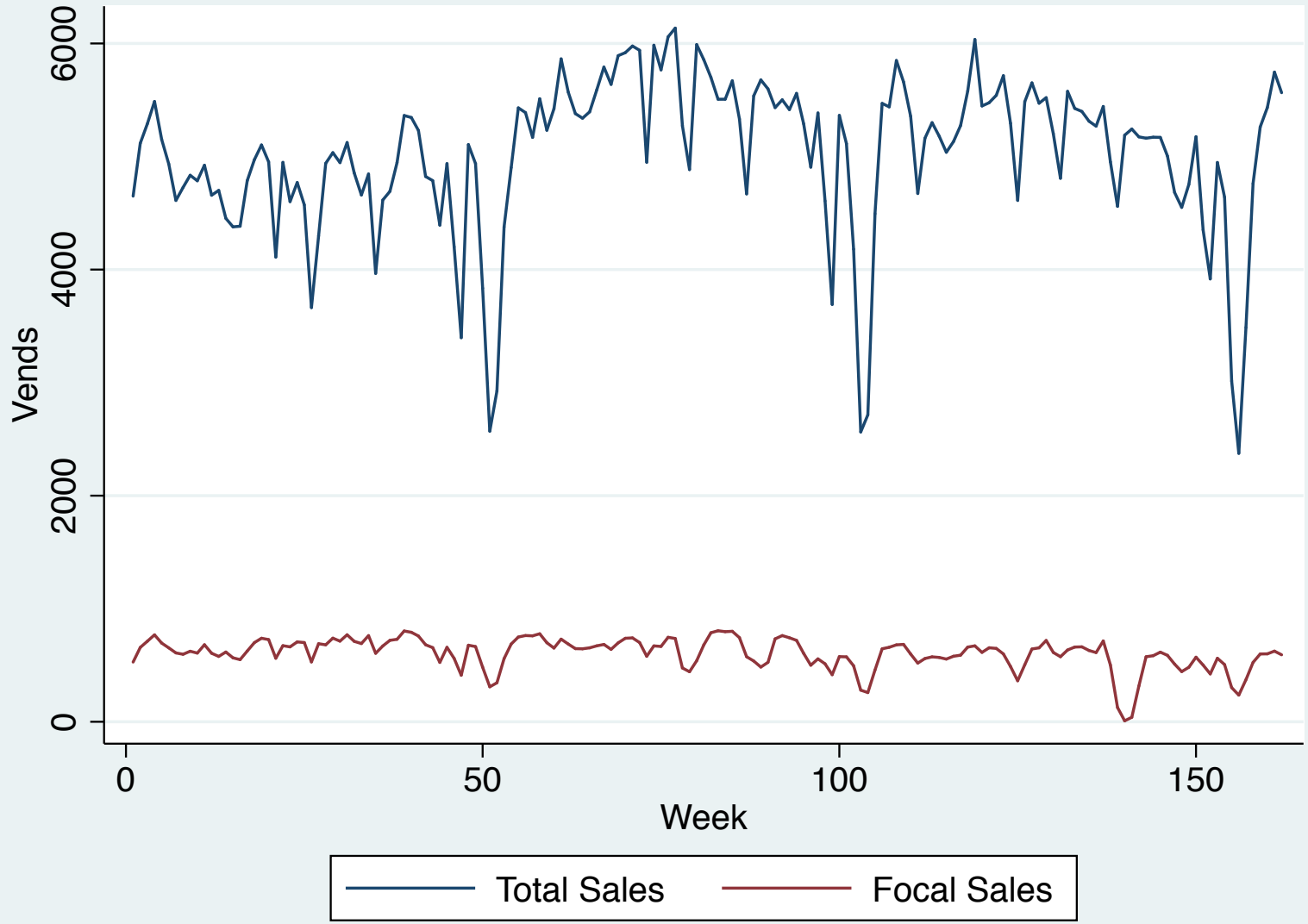
Here is an example where matching was helpful from a paper by Prof. Mortimer:

- She ran a randomized experiment where we removed Snickers bars from around 60 vending machines in office buildings in downtown Chicago.
- There are a few possible control groups:
  1. Same vending machine in other weeks (captures heterogeneous tastes in the cross section)
  2. Other vending machines in the same week (might capture aggregate shocks, ad campaigns, etc.)
- We went with #1 as #2 was not particularly helpful.

A Matching Example

Major problem was that there was a ton of heterogeneity in the overall level of (potential) weekly sales which we call M_t.

- Main source of heterogeneity is how many people are in the office that week, or how late they work.
- Based on total sales our average over treatment weeks was in the 74th percentile of all weeks.
- This was after removing a product, so we know sales should have gone down!
- How do we fix this without running the experiment for an entire year!


Vends
0
2000
4000
6000
0
50
100
150
Week
Total Sales
Focal Sales

A Matching Example

Ideally we could just observe M_t directly and use that as our matching variable X

- We didn't observe it directly and tried a few different measures:
  - Sales at the soda machine next to the snack machine
  - Sales of salty snacks at the same machine (not substitutes for candy bars).
  - We used k-NN with $k = 4$ to select control weeks – notice we re-weight so that overall sales are approximately same (minus the removed product).
- We also tried a more structured approach:
  - Define controls weeks as valid IFF
  - Overall sales were weakly lower
  - Overall sales were not less than Overall Sales less expected sales less Snickers Sales.

Product	Control Mean	Control %ile	Treatment Mean	Treatment %ile	Mean Difference
Vends					
Peanut M&Ms	359.9	73.6	478.3*	99.4	118.4*
Twix Caramel	187.6	55.3	297.1*	100.0	109.5*
Assorted Chocolate	334.8	66.7	398.0*	95.0	63.2*
Assorted Energy	571.9	63.5	616.2	76.7	44.3
Zoo Animal Cracker	209.1	78.6	243.7*	98.1	34.6*
Salted Peanuts	187.9	70.4	216.3*	93.7	28.4
Choc Chip Famous Amos	171.6	71.7	193.1*	95.0	21.5*
Ruger Vanilla Wafer	107.3	59.7	127.9	78.6	20.6*
Assorted Candy	215.8	43.4	229.6	60.4	13.7
Assorted Potato Chips	279.6	64.2	292.4*	66.7	12.8
Assorted Pretzels	548.3	87.4	557.7*	88.7	9.4
Raisinets	133.3	66.0	139.4	74.2	6.1
Cheetos	262.2	60.1	260.5	58.2	-1.8
Grandmas Choc Chip	77.9	51.3	72.5	37.8	-5.4
Doritos	215.4	54.1	203.1	39.6	-12.3*
Assorted Cookie	180.3	61.0	162.4	48.4	-17.9
Skittles	100.1	62.9	75.1*	30.2	-25.1*
Assorted Salty Snack	1382.8	56.0	1276.2*	23.3	-106.7*
Snickers	323.4	50.3	2.0*	1.3	-321.4*
Total	5849.6	74.2	5841.3	73.0	-8.3

Notes: Control weeks are selected through the-neighbor matching using four control observations for each treatment week. Percentiles are relative to the full distribution of control weeks.

How do you actually do this?

- One dimension is easy: just sort
- In multiple dimensions, there are a variety of built in nearest neighbor packages (Abadie Imbens (2006))
- What's nice about these is that the researcher only has to pick the number of matches (although the default tolerances not always innocuous)
- This is still cursed in that our nearest neighbors get further away as the dimension grows.
- Suppose instead we had a sufficient statistic

Propensity Score

- Rosenbaum and Rubin propose the propensity score

$$e(x) = Pr(W_i = 1|X_i) = E[W_i|X_i = x]$$

- They prove that under the assumption of unconfoundedness,

$$(Y_i(0), Y_i(1)) \perp W_i|e(X_i)$$

- So even if X is high dimensional, it is sufficient to condition on a scalar function
- Of course, the true propensity score is not known...

This suggests an attractive weigthing

4.B.3 Propensity Score Estimators: Weighting

$$\mathbb{E}\left[\frac{WY}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{WY_i(1)}{e(X)}\middle| X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{e(X)Y_i(1)}{e(X)}\right]\right] = \mathbb{E}[Y_i(1)],$$

and similarly

$$\mathbb{E}\left[\frac{(1-W)Y}{1-e(X)}\right] = \mathbb{E}[Y_i(0)],$$

implying

$$\tau_P = \mathbb{E}\left[\frac{W\cdot Y}{e(X)} - \frac{(1-W)\cdot Y}{1-e(X)}\right].$$

With the propensity score known one can directly implement this estimator as

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i\cdot Y_i}{e(X_i)} - \frac{(1-W_i)\cdot Y_i}{1-e(X_i)}\right). \tag{3}$$

Approaches now look similar

- One option is "inverse probability weighting"
- Nonparametrically estimate $e(x)$, the compute

$$\hat{\tau} = \sum_i^N \frac{W_i Y_i}{\hat{e}(X_i)} / \sum_i^N \frac{W_i}{\hat{e}(X_i)} - \sum_i^N \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)} / \sum_i^N \frac{(1-W_i)}{1-\hat{e}(X_i)}$$

  where this is slightly more complicated than just plugging in $\hat{e}()$ because in your sample the weights won't necessarily sum to one (Hirano, Imbens and Ridder (2003))
- Alternatively we could flexibly estimate μ_w then plug in these predictions for each observation manually.
- With discrete covariates, these will be equivalent!
- Otherwise there finite sample propoerties will vary depending on the smoothness of the regression and propensity score functions.

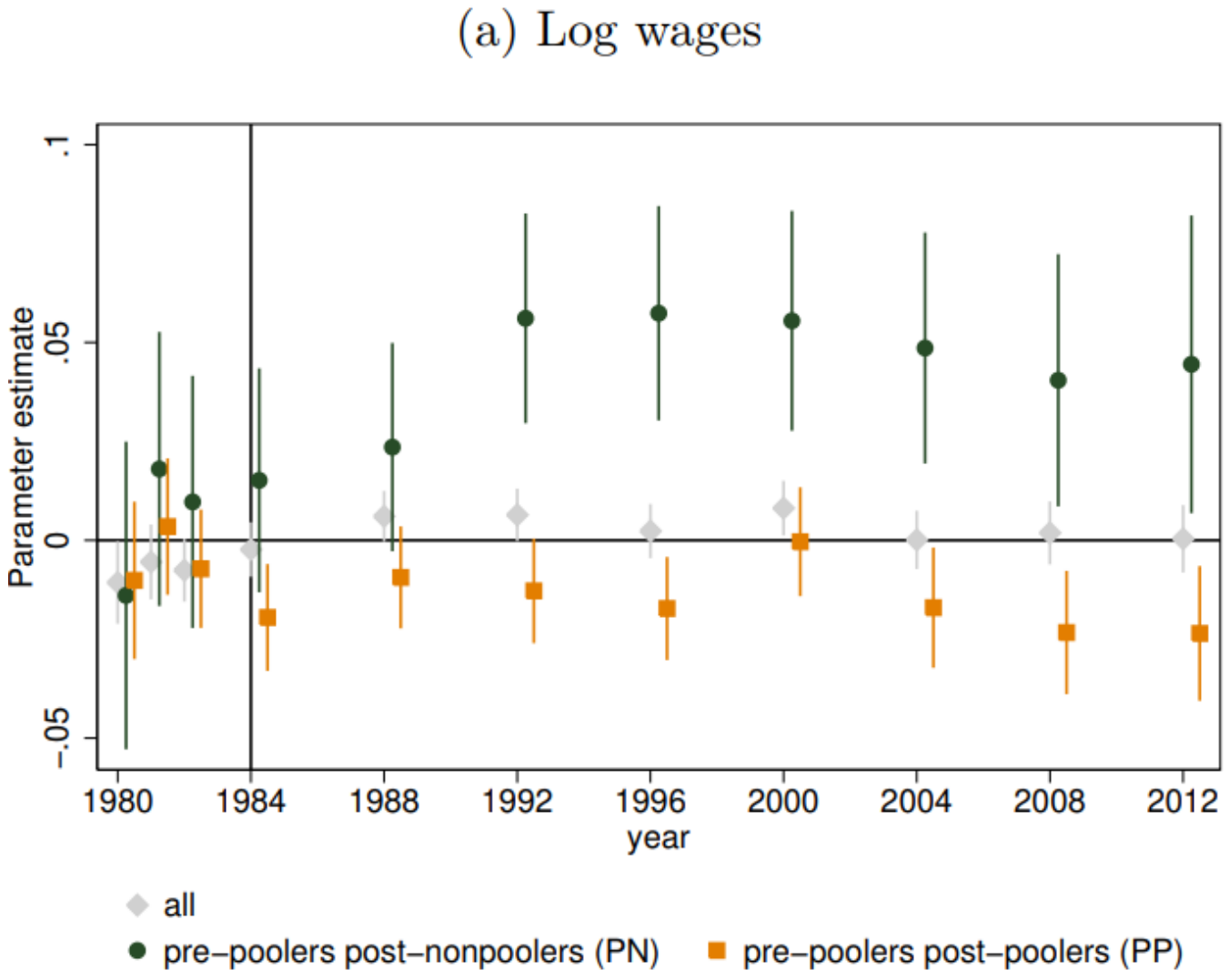
What about matching on the (estimated) propensity score?

- VERY widely used approach
- Large sample properties not known
- "Why Propensity Scores Should Not Be Used for Matching" (King and Nielsen, Forthcoming)
- Show this performs poorly in simulations compared to matching on X's directly.
- One alternative from the same author's: Coarsened Exact Matching
  - Available in R and Stata from Gary King's website
  - The idea: temporarily coarsen each variable into substantively meaningful groups, exact match on these coarsened data, and then retain only the original (uncoarsened) values of the matched data.

CEM has many uses

- Linh To's JMP:
- Question: Is there a signal value to parental leave?
- Theory: many PBNE's. In practice depends on pooling.
- Setting: Extension of leave in Denmark.
- Look for response among three types of women:
  1. pool, pool
  2. pool, separate
  3. separate, separate
- Convincing RD: restrict to sample already pregnant when law announced
- Challenge: Only see mothers in one group or the other
- Solution: Match each pre period mother using their closest post-period counterpart, and assign her to that post-group.

Results


(a) Log wages
Parameter estimate
.1
.05
0
−.05
1980
1984
1988
1992
1996
2000
2004
2008
2012
year
all
pre−poolers post−nonpoolers (PN)
pre−poolers post−poolers (PP)

What can ML add here?

- Estimating the propensity score is a pure **prediction** problem. We don't care what causes someone to be treated in this setup
- This is a natural place for ML (decision trees, random forests).
- What should we use to predict?

Some recent ML proposals I

Belloni, Chernozhukov, FernÃ¡ndez, and Hansen (2013)

- "double selection" procedure
- use LASSO to select X which predict Y, and another LASSO to find X that predict W
- then do OLS on the union of the two sets of covariates
- show this performs better than simple regularized regression of outcome on treatment and covariates in one step

Some recent ML proposals II

Athey, Imbens, and Wager (2016)
"Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions)"

- Idea: In order to predict the counterfactual outcomes that the treatment group would have had in the absence of the treatment, it is necessary to extrapolate from control
- This is confounded by imbalance.
- AIW construct weights so these samples are equivalent, and run penalized regression to compute τ

Assessing Unconfoundedness

- This assumption is fundamentally untestable
- However people have proposed a number of tests which, if failed, might be *inconsistent* with unconfoundedness.
- One option is to look for an "effect" on an untreated group.
- Imagine you had one sample of "eligible" units, some who were treated and some who weren't. And another sample of "ineligible" units, all of whom are also untreated by construction.
- You could estimate a difference in outcomes within the two untreated groups. If eligible but untreated units look different than uneligible, that should be worrisome.
- Imbens lecture does this with the Lalonde data and the CPS.
- Another natural approach is to use "psuedo outcomes", like lagged Y.

Assessing Overlap

- Obviously want to start with a summary table comparing the means of your treatment and control groups.
- What's a big difference? t-stats reflective of sample size
- Instead report the normalized difference in covariates. According to Imbens, a an average difference bigger than 0.25 standard deviations is worrisome.
- Another alternative is to plot the propensity score for the two groups.

Matching wrapup

- Even under unconfoundedness, very important to ensure overlap
- Restrict your sample so that its balanced, using exact matching if low dimensional, coarse or propensity score otherwise
- Assess unconfoundedness using a psuedo-outcome if possible
- Run regression on your matched sample

Important: Do not match on outcomes!

- Controlling for covariates increases the likelihood of a causal regression estimate, but more controls are not always better.
- Let's say your treatment is in fact randomly assigned. But it affects multiple outcomes. In this case, controlling for these other outcomes can actually make things worse!
- MHE call this the "bad control" problem (p 64)

MHE college / occupation example

- W_i denotes white collar workers, Y_i denotes earnings.
- Both determined by college graduation status C_i

$$\begin{aligned} Y_i &= C_i Y_{1i} + (1 - C_i) Y_{0i} \\ W_i &= C_i W_{1i} + (1 - C_i) W_{0i} \end{aligned}$$

- Assume that C_i randomly assigned, so independent of all outcomes. We can easily estimate the causal effect on either:

$$\begin{aligned} E[Y_i|C_i = 1] - E[Y_i|C_i = 0] &= E[Y_{1i} - Y_{0i}] \\ E[W_i|C_i = 1] - E[W_i|C_i = 0] &= E[W_{1i} - W_{0i}] \end{aligned}$$

What if we estimate the impact of college *conditional on a white collar job*?

- Condition on $W_i = 1$

$$E[Y_i|W_i = 1, C_i = 1] - E[Y_i|W_i = 1, C_i = 0]$$
$$= E[Y_{1i}|W_{1i} = 1, C_i = 1] - E[Y_{0i}|W_{0i} = 1, C_i = 0]$$

- By joint independence of outcomes and C, this is

$$E[Y_{1i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1]$$

What if we estimate the impact of college *conditional on a white collar job*?

- This can be decomposed into:

$$\begin{aligned} &\text{Casual effect} && E[Y_{1i} - Y_{0i}|W_{1i}] + \\ &\text{Selection bias} && E[Y_{0i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1] \end{aligned}$$

- in words: we have the causal effect, plus the fact that college changes the composition of the pool of white collar workers
- bias can go either way: but point is that it is there even if there is no causal impact of college on wages.

Instrumental Varibales

See Guido Imben's NBER Slides.

How Close to ATE?

Angrist and Imbens give some idea how close to the ATE the LATE is:

$$\widehat{\beta}_1^{TSLS} \rightarrow^p \frac{E[\beta_{1i}\pi_{1i}]}{E[\pi_{i1}]} = LATE$$

$$LATE = ATE + \frac{Cov(\beta_{1i}, \pi_{1i})}{E[\pi_{1i}]}$$

- Average TE weighted by the probability that each individual's treatment is influenced by Z_i.
- If you always (never) get treated you don't show up in LATE.

How Close to ATE?

- With different instruments you get different π_{1i} and TSLS estimators!
- Even with two valid Z_1, Z_2
  - Can be influential for different members of the population.
  - Using Z_1, TSLS will estimate the treatment effect for people whose probability of treatment X is most influenced by Z_1
  - The LATE for Z_1 might differ from the LATE for Z_2

Example: Cardiac Catheterization

- $Y_i =$ surival time (days) for AMI patients
- $X_i =$ whether patient received cadiac catheterization (or not) (intensive treatment)
- $Z_i =$ differential distance to CC hospital

$$\begin{aligned} SurvivalDays_i &= \beta_0 + \beta_{1i} CardCath_i + u_i \\ CardCath_i &= \pi_0 + \pi_{1i} Distance_i + v_i \end{aligned}$$

- For whom does distance have the great effect on probability of treatment?
- For those patients what is their β_{1i}?

Example: Cardiac Catheterization

- IV estimates causal effect for patients whose value of X_i is most heavily influenced by Z_i
  - Patients with small positive benefit from CC in the expert judgement of EMT will receive CC if trip to CC hospital is short (compliers)
  - Patients that need CC to survive will always get it (always-takers)
  - Patients for which CC would be unnecessarily risky or harmful will not receive it (never-takers)
  - Patients for who would have gotten CC if they lived further from CC hospital (hopefully don't see) (defiers)
- We mostly weight towards the people with small positive benefits.

Local Average Treatment Effect

So how is this useful?

- It shows why IV can be meaningless when effects are heterogeneous.
- It shows that if the monotonicity assumption can be justified, IV estimates the effect for a particular subset of the population.
- In general the estimates are specific to that instrument and are not generalisable to other contexts.
- As an example consider two alternative policies that can increase participation in higher education.
  - Free tuition is randomly allocated to young people to attend college ($Z_1 = 1$ means that the subsidy is available).
  - The possibility of a competitive scholarship is available for free tuition ($Z_1 = 1$ means that the individual is allowed to compete for the scholarship).

Local Average Treatment Effect

- Suppose the aim is to use these two policies to estimate the returns to college education. In this case, the pair $\{Y^1, Y^0\}$ are log earnings, the treatment is going to college, and the instrument is one of the two randomly allocated programs.
- First, we need to assume that no one who intended to go to college will be discouraged from doing so as a result of the policy (monotonicity).
- This could fail as a result of a General Equilibrium response of the policy; for example, if it is perceived that the returns to college decline as a result of the increased supply, those with better outside opportunities may drop out.

Local Average Treatment Effect

- Now compare the two instruments.
- The subsidy is likely to draw poorer liquidity constrained students into college but not necessarily those with the highest returns.
- The scholarship is likely to draw in the best students, who may also have higher returns.
- It is not a priori possible to believe that the two policies will identify the same parameter, or that one experiment will allow us to learn about the returns for a broader/different group of individuals.

Example: Pretrial Detention

- In US, innocent until proven guilty.
- Some defendants are detained prior to trial.
- Extreme cases are obvious, but lots of discretion in the middle.
- What are the impacts on:
  - time served
  - future crime
  - rehabilitation in to workforce

Example: Pretrial Detention

The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges†

By Will Dobbie, Jacob Goldin, and Crystal S. Yang*

Over 20 percent of prison and jail inmates in the United States are currently awaiting trial, but little is known about the impact of pretrial detention on defendants. This paper uses the detention tendencies of quasi-randomly assigned bail judges to estimate the causal effects of pretrial detention on subsequent defendant outcomes. Using data from administrative court and tax records, we find that pretrial detention significantly increases the probability of conviction, primarily through an increase in guilty pleas. Pretrial detention has no net effect on future crime, but decreases formal sector employment and the receipt of employment- and tax-related government benefits. These results are consistent with (i) pretrial detention weakening defendants' bargaining positions during plea negotiations and (ii) a criminal conviction lowering defendants' prospects in the formal labor market. (*JEL* J23, J31, J65, K41, K42)

Means for detained vs released defendants

Panel E. Outcomes		
Any guilty offense	0.578	0.486
Guilty plea	0.441	0.207
Any incarceration	0.300	0.145
Failure to appear in court	0.121	0.179
Rearrest in 0–2 years	0.462	0.398
Earnings ($ thousands) in 1–2 years	5.224	7.911
Employed in 1–2 years	0.378	0.509
Any income in 1–2 years	0.458	0.522
Earnings ($ thousands) in 3–4 years	5.887	8.381
Employed in 3–4 years	0.378	0.483
Any income in 3–4 years	0.461	0.508
Observations	186,938	234,127

Notes: This table reports descriptive statistics for the sample of defendants from Philadelphia and Miami-Dade counties. Data from Philadelphia are from 2007–2014 and data from Miami-Dade are from 2006–2014. Information on ethnicity, gender, age, and criminal outcomes is derived from court records. Information on earnings, employment, and income is derived from the IRS data and is only available for the 77 percent of the criminal records matched to these data. See the online data Appendix for additional details on the sample and variable

First stage: Judges Matter

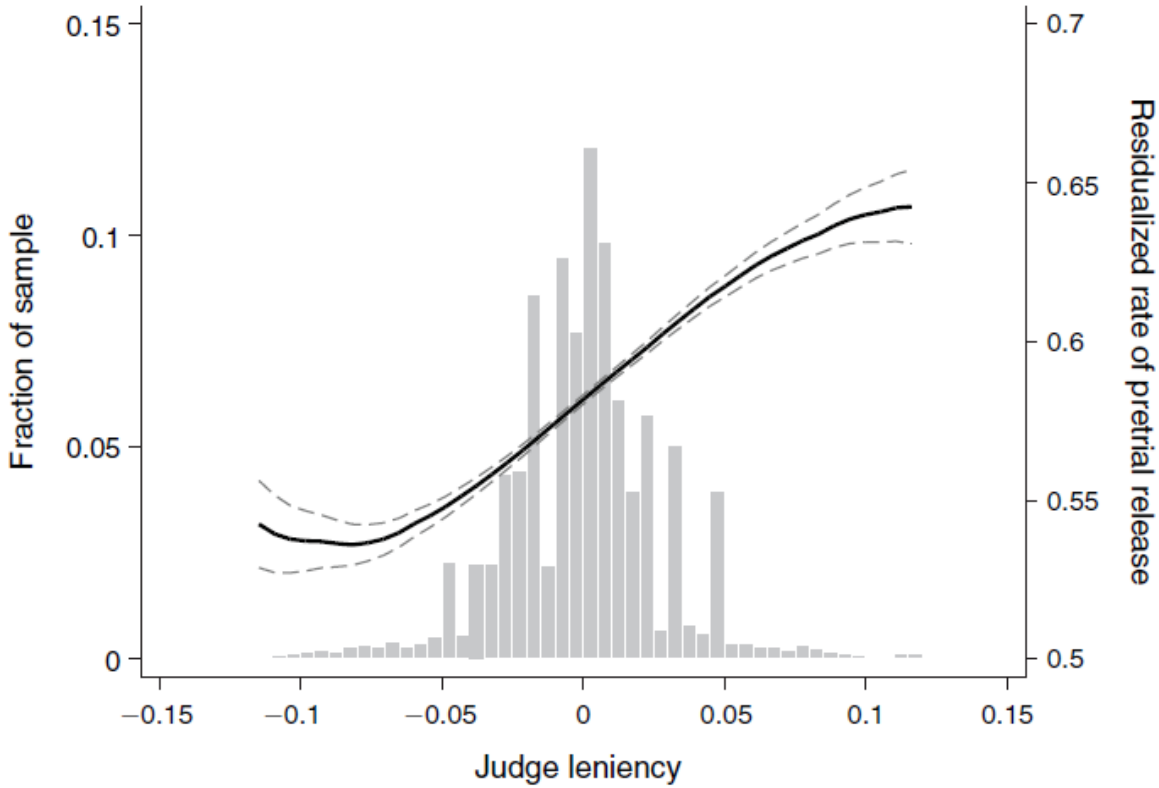


FIGURE 1. DISTRIBUTION OF JUDGE LENIENCY MEASURE AND FIRST STAGE

Note: This figure reports the distribution of the judge leniency measure that is estimated using data from other cases assigned to a bail judge in the same year following the procedure described in Section III.

Is assignment random?

TABLE 3—TEST OF RANDOMIZATION

	Pretrial release (1)	Judge leniency (2)
Male	−0.11781	0.00007
	(0.00716)	(0.00015)
Black	−0.03941	0.00003
	(0.00362)	(0.00017)
Age at bail decision	−0.01287	−0.00005
	(0.00236)	(0.00006)
Prior offense in past year	−0.15492	0.00019
	(0.00739)	(0.00012)
Number of offenses	−0.02409	0.00000
	(0.00120)	(0.00002)
Felony offense	−0.25575	0.00005
	(0.01821)	(0.00010)
Any drug offense	0.12528	0.00013
	(0.00909)	(0.00019)
Any DUI offense	0.10966	0.00019
	(0.01679)	(0.00024)

Results

TABLE 4—PRETRIAL RELEASE AND CRIMINAL OUTCOMES

	Detained mean	OLS results			2SLS results	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Case outcomes						
Any guilty offense	0.578	−0.072	−0.057	−0.046	−0.123	−0.140
	(0.494)	(0.014)	(0.009)	(0.007)	(0.047)	(0.042)
Guilty plea	0.441	−0.188	−0.099	−0.082	−0.095	−0.108
	(0.497)	(0.008)	(0.010)	(0.007)	(0.056)	(0.052)
Any incarceration	0.300	−0.161	−0.104	−0.110	0.006	−0.012
	(0.458)	(0.012)	(0.006)	(0.007)	(0.029)	(0.030)
Panel B. Court process outcomes						
Failure to appear in court	0.121	0.063	0.010	0.021	0.158	0.156
	(0.326)	(0.004)	(0.008)	(0.007)	(0.046)	(0.046)
Absconded	0.002	0.005	0.002	0.002	0.005	0.005
	(0.045)	(0.000)	(0.000)	(0.000)	(0.004)	(0.004)
Panel C. Future crime						
Rearrest in 0–2 years	0.462	−0.050	−0.015	0.016	0.024	0.015
	(0.499)	(0.011)	(0.006)	(0.005)	(0.061)	(0.063)
Rearrest prior to disposition	0.155	0.051	0.066	0.100	0.192	0.189
	(0.362)	(0.008)	(0.007)	(0.007)	(0.038)	(0.042)
Rearrest after disposition	0.343	−0.075	−0.049	−0.041	−0.114	−0.121
	(0.475)	(0.006)	(0.002)	(0.003)	(0.057)	(0.055)
Court × time fixed effects	—	Yes	Yes	Yes	Yes	Yes
Baseline controls	—	No	Yes	Yes	No	Yes
Complier weights	—	No	No	Yes	No	No
Observations	186,938	421,065	421,065	421,065	421,065	421,065

Notes This table reports OLS and two-stage least squares results of the impact of pre-trial release. The regressions are estimated on the sample as described in the notes to Table 1. The dependent variable is listed in each

Interpretation: Who is marginal here?

Interpretation: Who is marginal here?

- Instrument isn't binary here
- Thought experiment is the same though: identify which defendants get out under the most lenient judge minus those that get out under the strictest judge

Table C.1: Sample Share by Compliance Type

Model Specification:	Local Linear Model			Linear Model		
Leniency Cutoff:	1%	1.5%	2%	1%	1.5%	2%
Compliers	0.13	0.13	0.13	0.11	0.10	0.09
Never Takers	0.36	0.36	0.36	0.39	0.39	0.40
Always Takers	0.51	0.51	0.51	0.50	0.51	0.51

Who are the compliers?

- Follow strategy of Dahl et al (QJE 2014)
- Estimate complier share by subgroup

Table C.2: Characteristics of Marginal Defendants

	$P[X=x]$	$P[X=x\|\text{complier}]$	$\frac{P[X=x\|\text{complier}]}{P[X=x]}$
White	0.402	0.375	0.931
	(0.001)	(0.017)	(0.042)
Non-White	0.598	0.624	1.047
	(0.001)	(0.017)	(0.028)
Drug	0.274	0.301	1.099
	(0.001)	(0.015)	(0.054)
Non-Drug	0.726	0.699	0.963
	(0.001)	(0.015)	(0.020)
Violent	0.173	0.010	0.058
	(0.001)	(0.012)	(0.068)
Non-Violent	0.827	0.990	1.197
	(0.001)	(0.012)	(0.014)
Felony	0.459	0.318	0.692
	(0.001)	(0.016)	(0.036)
Misdemeanor	0.541	0.682	1.261
	(0.001)	(0.016)	(0.030)
Prior Last Year	0.269	0.310	1.154
	(0.001)	(0.013)	(0.049)
No Prior	0.731	0.690	0.943
	(0.001)	(0.013)	(0.018)
Employed	0.475	0.457	0.963
	(0.001)	(0.017)	(0.036)
Non-Employed	0.525	0.543	1.033
	(0.001)	(0.017)	(0.033)

How useful is LATE here?

- What can you do with this estimate?
- Is it of policy importance?

Example: Dams

DAMS*

ESTHER DUFLO AND ROHINI PANDE

This paper studies the productivity and distributional effects of large irrigation dams in India. Our instrumental variable estimates exploit the fact that river gradient affects a district's suitability for dams. In districts located downstream from a dam, agricultural production increases, and vulnerability to rainfall shocks declines. In contrast, agricultural production shows an insignificant increase in the district where the dam is located but its volatility increases. Rural poverty declines in downstream districts but increases in the district where the dam is built, suggesting that neither markets nor state institutions have alleviated the adverse distributional impacts of dam construction.

I. INTRODUCTION

"If you are to suffer, you should suffer in the interest of the country." Indian Prime Minister Nehru, speaking to those displaced by Hirakud Dam, 1948.

Example: Dams

- What is the exclusion restriction here?
- How useful is this LATE?

Weak instruments

- So far we have assumed that the instrument is relevant

$$cov(Z, W) > 0$$

- Intuitively, if there are no "compliers", we can't learn anything from IV.
- In applications, instruments are sometimes barely relevant, i.e. $\hat{C}ov(dz, x) \neq 0$, but it's close.
- This leads to:
  - Large finite sample bias of $\hat{\beta}^{2SLS}$
  - Inference issues: (wrong standard error, incorrect p-values, incorrect confidence intervals)

Setup:

$$Y_i = X_i\beta + \varepsilon_i \quad \text{(Structural equation)} \quad (1)$$

$$X_i = Z_i'\pi + V_i \quad \text{(First stage)} \quad (2)$$

$$Y_i = Z_i'\delta + U_i, \quad \delta = \pi\beta, \varepsilon = U - \beta V. \quad \text{(Reduced form)} \quad (3)$$

The two conditions for instrument validity

(i) Relevance: $\text{cov}(Z,X) \neq 0$ or $\pi \neq 0$ (general k)

(ii) Exogeneity: $\text{cov}(Z,\varepsilon) = 0$

The IV estimator when $k = 1$ (Wright 1926)

$$\text{cov}(Z,Y) = \text{cov}(Z, X\beta + \varepsilon) = \text{cov}(Z,X)\beta + \text{cov}(Z,\varepsilon)$$
$$= \text{cov}(Z,X)\beta \quad \text{by (i)}$$

so

$$\beta = \frac{\text{cov}(Z,Y)}{\text{cov}(Z,X)} \quad \text{by (ii)}$$

IV estimator:

$$\hat{\beta}^{IV} = \frac{n^{-1}\sum_{i=1}^{n} Z_i Y_i}{n^{-1}\sum_{i=1}^{n} Z_i X_i} = \frac{\hat{\delta}}{\hat{\pi}}$$

Setup:

$$Y_i = X_i\beta + \varepsilon_i \quad \text{(Structural equation)} \quad (1)$$

$$X_i = Z_i'\pi + V_i \quad \text{(First stage)} \quad (2)$$

$$Y_i = Z_i'\delta + U_i, \quad \delta = \pi\beta, \varepsilon = U - \beta V. \quad \text{(Reduced form)} \quad (3)$$

$k > 1$: Two stage least squares (TSLS)

$$\hat{\beta}^{TSLS} = \frac{n^{-1}\sum_{i=1}^{n}\hat{X}_i Y_i}{n^{-1}\sum_{i=1}^{n}\hat{X}_i^2}, \quad \text{where } \hat{X}_i = \text{predicted value from first stage}$$

$$= \frac{\mathbf{X'Z(Z'Z)^{-1}Z'Y}}{\mathbf{X'Z(Z'Z)^{-1}Z'X}}$$

$$= \frac{\hat{\pi}'\hat{Q}_{ZZ}\hat{\delta}}{\hat{\pi}'\hat{Q}_{ZZ}\hat{\pi}}, \quad \text{where } \hat{Q}_{ZZ} = n^{-1}\sum_{i=1}^{n} Z_i Z_i'$$

The weak instruments problem is a "divide by zero" problem

- $cov(Z,X)$ is nearly zero; or π is nearly zero; or
- $\hat{\pi}'\hat{Q}_{ZZ}\hat{\pi}$ is noisy
- Weak IV is a subset of weak identification (Stock-Wright 2000, Nelson-Starts 2006, Andrews-Cheng 2012)

Weak instruments

- This is an active area of research. See Angrist and Pischke (Ch. 4); or Stock and Andrews 2018 NBER minicourse for a recent treatment.
- Always report first stage F statistic for significance of coefficients on instruments - rule of thumb: $F \geq 10$ is okay (under weak instrument asymptotics, bias of 2SLS and is $< 10\%$ when $F \geq 10$.)
- In general, adding weak instruments makes it worse!
- Estimates approach OLS. If instrument doesn't satisfy exclusion restriction, this could be even worse!

LASSO for selecting instruments

- Data often gives us many plausibly relevant instruments that satisfy the exclusion restriction. Which should we use?
- We know that adding many weak instruments is problematic.
- Intuitively we want something this is highly *predictive* of the endogenous variable. This is what Lasso is good at. (Belloni et al., 2012)

Application: Eminent Domain

- How do changes in the government's ability to appropriate property affect property markets?
- Challenge: Changes likely endogenous to the strength of those markets and other economic factors
- Even if law changes are endogenous, much of the real world variation comes from court rulings.
- Instrument: Judges

IV Challenge: Which judges are more inclined to rule for/ against eminent domain?

- Unlike pretrial detention example, don't have large N of other cases.
- Many judge characteristics: gender, race, religion, political affiliation, whether the judge's bachelor's degree was obtained in-state, whether the bachelor's degree is from a public university, whether the JD was obtained from a public university, and whether the judge was elevated from a district court.
- All are randomly assigned. Which ones are *relevant*?

How do we typically proceed here?

- Pick the ones that make the most sense on intuitive grounds.
- In another paper, Chen and Yeh do exactly this, using
  1. whether a judge did not report a religious affiliation
  2. whether the judge earned her law degree from a public institution
- Could try other instruments and see if results are "robust" (should they be?)
- Could try everything: data mining/ not feasible
- Belloni et al. create 140 first stage vars, and let LASSO decide.
- Since all satisfy the exclusion restriction (by assumption), this first stage selection has no bearing on second stage interpretation.

Results

EFFECT OF FEDERAL APPELLATE TAKINGS LAW DECISIONS ON ECONOMIC OUTCOMES[a]

	Home Prices			GDP
	log(FHFA)	log(Non-Metro)	log(Case–Shiller)	log(GDP)
Sample Size	312	110	183	312
OLS	0.0114	0.0108	0.0152	0.0099
s.e.	0.0132	0.0066	0.0132	0.0048
2SLS	0.0262	0.0480	0.0604	0.0165
s.e.	0.0441	0.0212	0.0296	0.0162
FS-W	28.0859	82.9647	67.7452	28.0859
Post-LASSO	0.0369	0.0357	0.0631	0.0133
s.e.	0.0465	0.0132	0.0249	0.0161
FS-W	44.5337	243.1946	89.5950	44.5337
S	1	4	2	1
Post-LASSO+	0.0314	0.0348	0.0628	0.0144
s.e.	0.0366	0.0127	0.0245	0.0131
FS-W	73.3010	260.9823	105.3206	73.3010
S	3	6	3	3
Spec. Test	−0.2064	0.5753	−0.0985	0.1754

[a]This table reports the estimated effect of an additional pro-plaintiff takings decision, a decision that goes against the government and leaves the property in the hands of the private owner, on various economic outcomes using two-stage least squares (2SLS). The characteristics of randomly assigned judges serving on the panel that decides the case are used as instruments for the decision variable. All estimates include circuit effects, circuit-specific time trends, time effects, controls for the number of cases in each circuit-year, and controls for the demographics of judges available within each circuit-year. Each column corresponds to a different dependent variable. log(FHFA), log(Non-Metro), and log(Case–Shiller) are within-circuit averages of log-house-price-indexes, and log(GDP) is the within-circuit average of log of state-level GDP. OLS are ordinary least squares estimates. 2SLS is the 2SLS estimator with the original

Regression Discontinuity Design

- Another popular research design is the Regression Discontinuity Design.
- In some sense this is a special case of IV regression. (RDD estimates a LATE).
- Most of Chris's slides taken from Lee and Lemieux (2010).
- For an extensive recent treatment, see "A Practical Introduction to Regression Discontinuity Designs" (Cattaneo, Idrobo and Titiunik (2019, CUP)) (available here)
- Matias Catteneo has a number of useful tools (in R and Stata) available on his website.

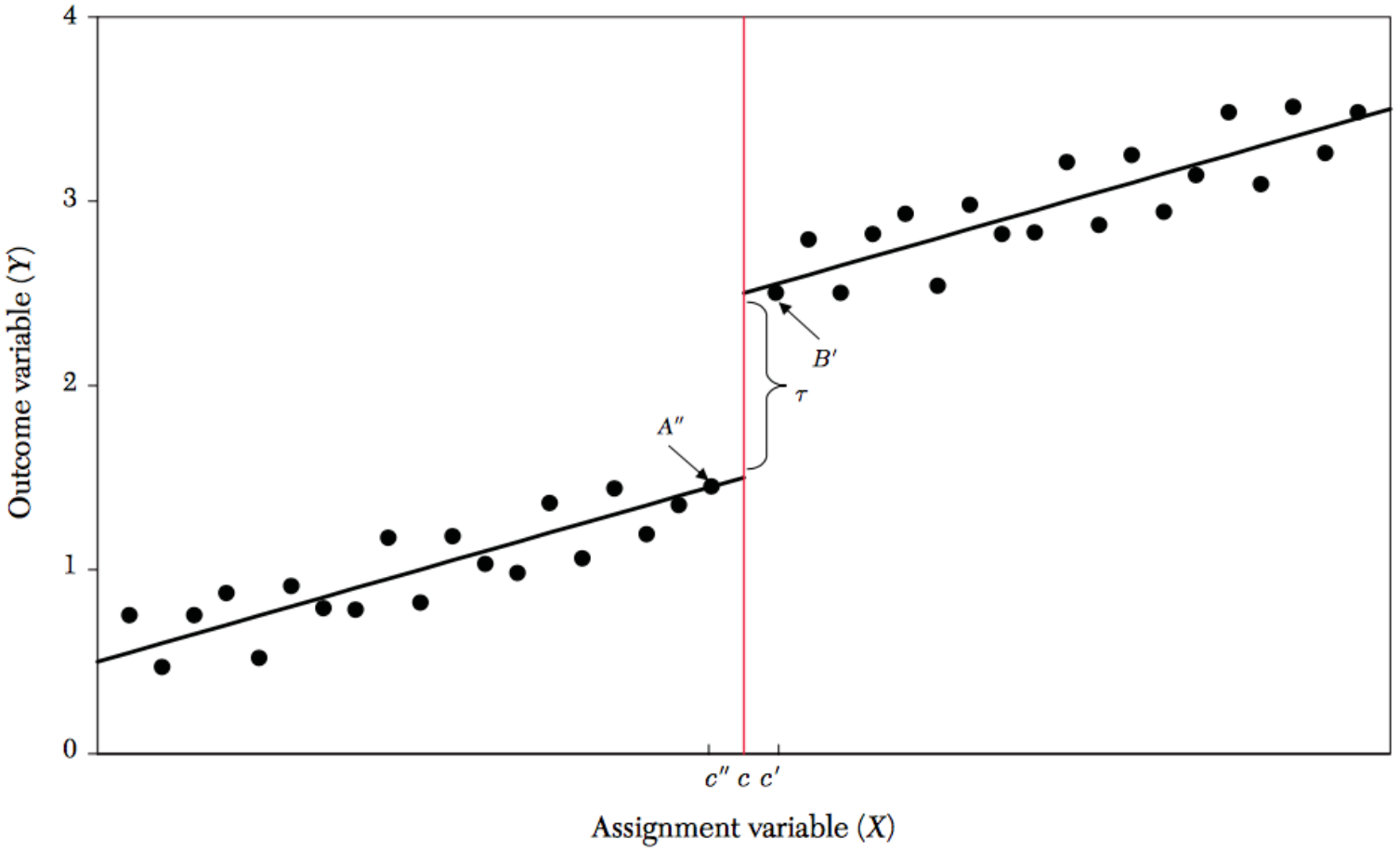
RDD: Basics

- We have a running or forcing variable x such that

$$\lim_{x \to c^+} P(T_i|X_i = x) \neq \lim_{x \to c^-} P(T_i|X_i = x)$$

- The idea is that there is a discontinuous jump in the probability of being treated.
- For now we focus on the sharp discontinuity: $P(T_i|X_i \geq c) = 1$ and $P(T_i|X_i < c) = 0$
- There is no single x for which we observe treatment and control. (Compare to Propensity Score!).
- The most important assumption is that of no manipulability $\tau_i \perp D_i$ in some neighborhood of c.
- Example: a social program is available to people who earned less than $25,000.
  - If we could compare people earning $24,999 to people earning $25,001 we would have as-if random assignment. (MAYBE)
  - But we might not have that many people...

RDD: In Pictures



RDD: Sharp RD Case

RDD uses a set of assumptions distinct from our LATE/IV assumptions. Instead it depends on continuity.

- We need that $E[Y^{(1)}|X]$ and $E[Y^{(0)}|X]$ both be continuous at $X = c$.
- People just to the left of c are a valid control for those just to the right of c.
- This is not a testable assumption
  - Typically draw pictures of *other* X's at c
- Most basic approach is regression

$$Y_i = \beta_0 + \tau D_i + X_i\beta + \epsilon_i$$

where $D_i = \mathbf{1}[X_i > c]$

- This puts a lot of restrictions (linearity) on the relationship between Y and X.

RDD: Nonlinearity

First thing to relax is assumption of linearity.

$$Y_i = f(x_i) + \tau D_i + \epsilon_i$$

- Two options for $f(x_i)$:
  1. Kernels: Local Linear Regression
  2. Polynomials:
     $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x^p + \tau D_i + \epsilon_i.$
     - Actually, people suggest different polynomials on each side of cutoff! (Interact everything with D_i).
- Same objective. Want to flexibly capture what happens on both sides of cutoff.
- Otherwise risk confusing nonlinearity with discontinuity!

RDD: Kernel Boundary Problem

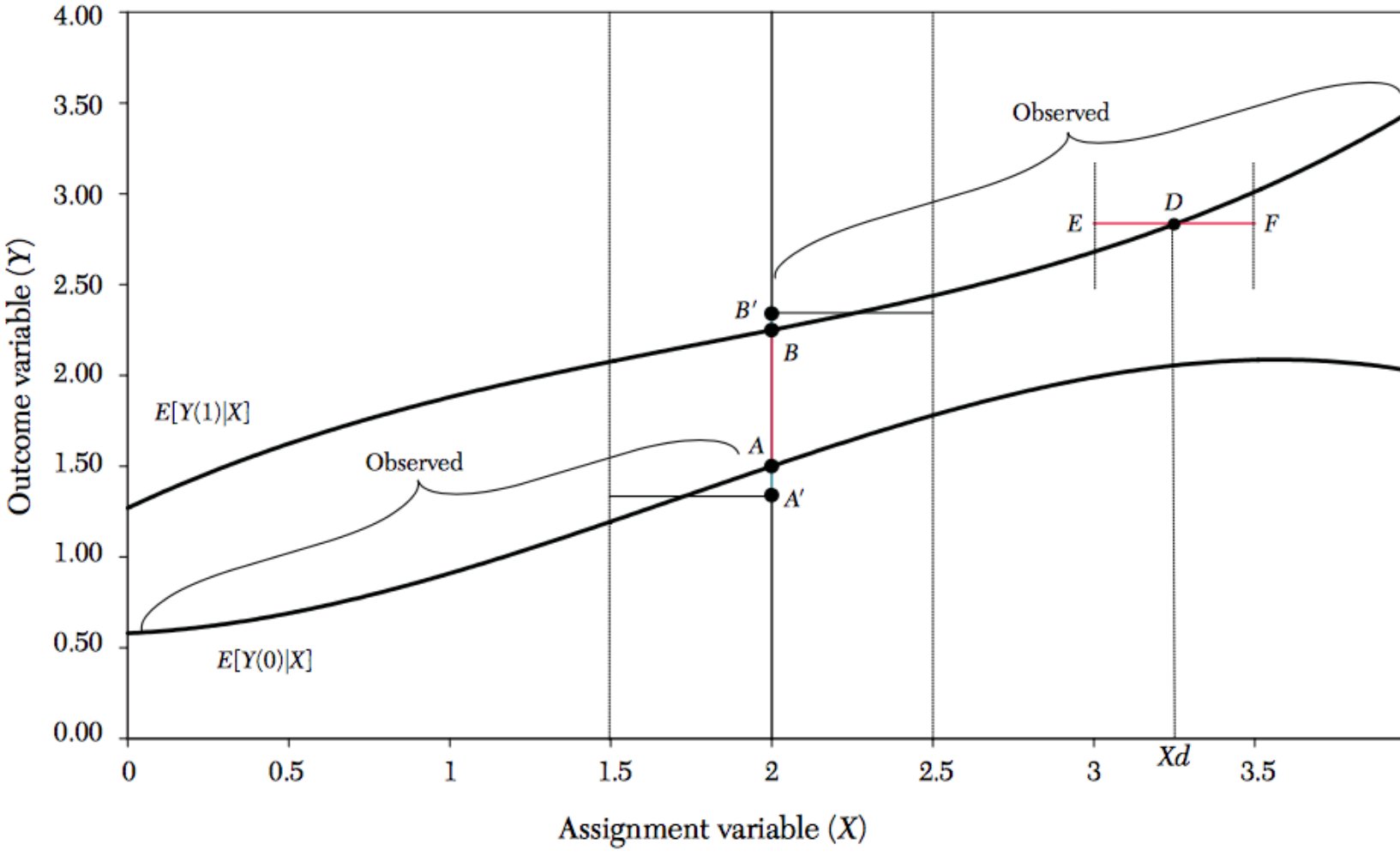
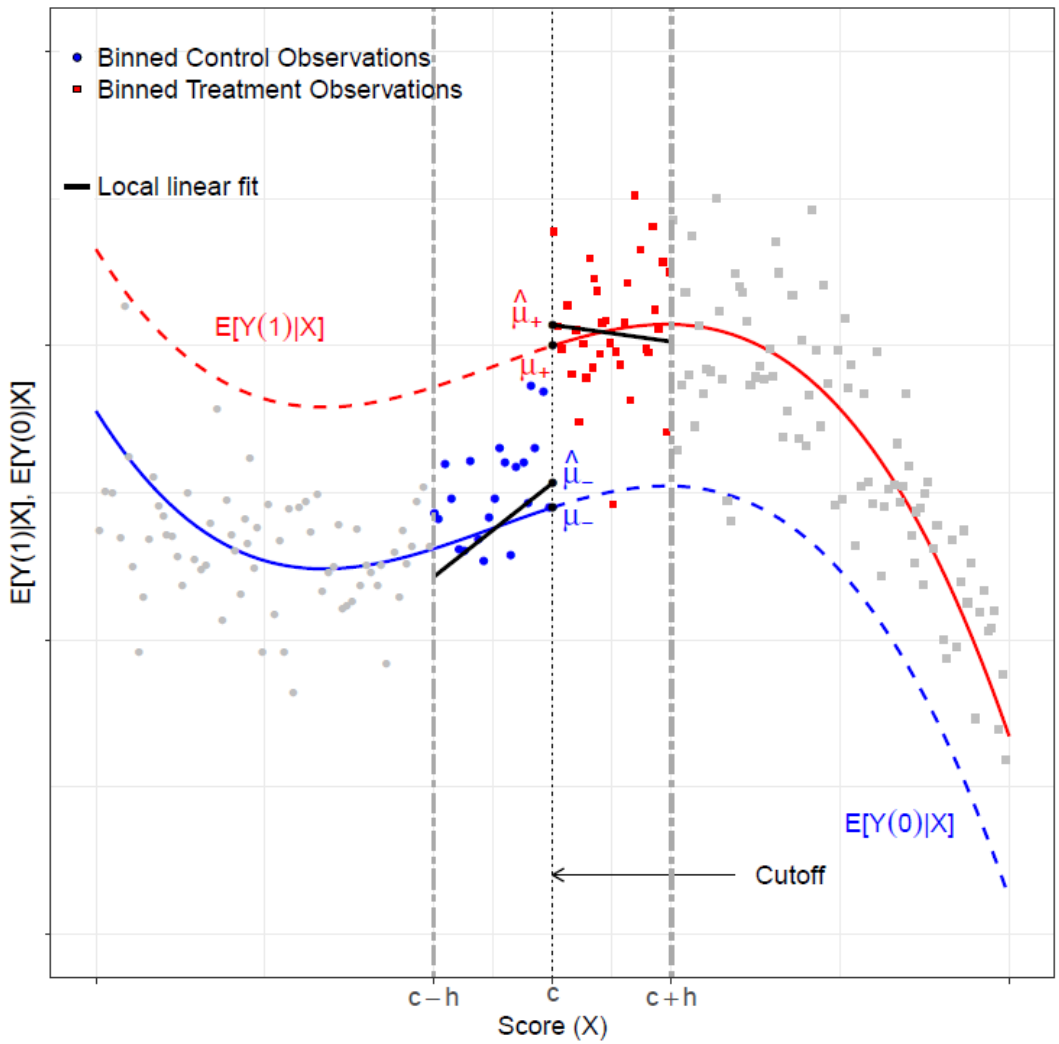


Figure 2. Nonlinear RD

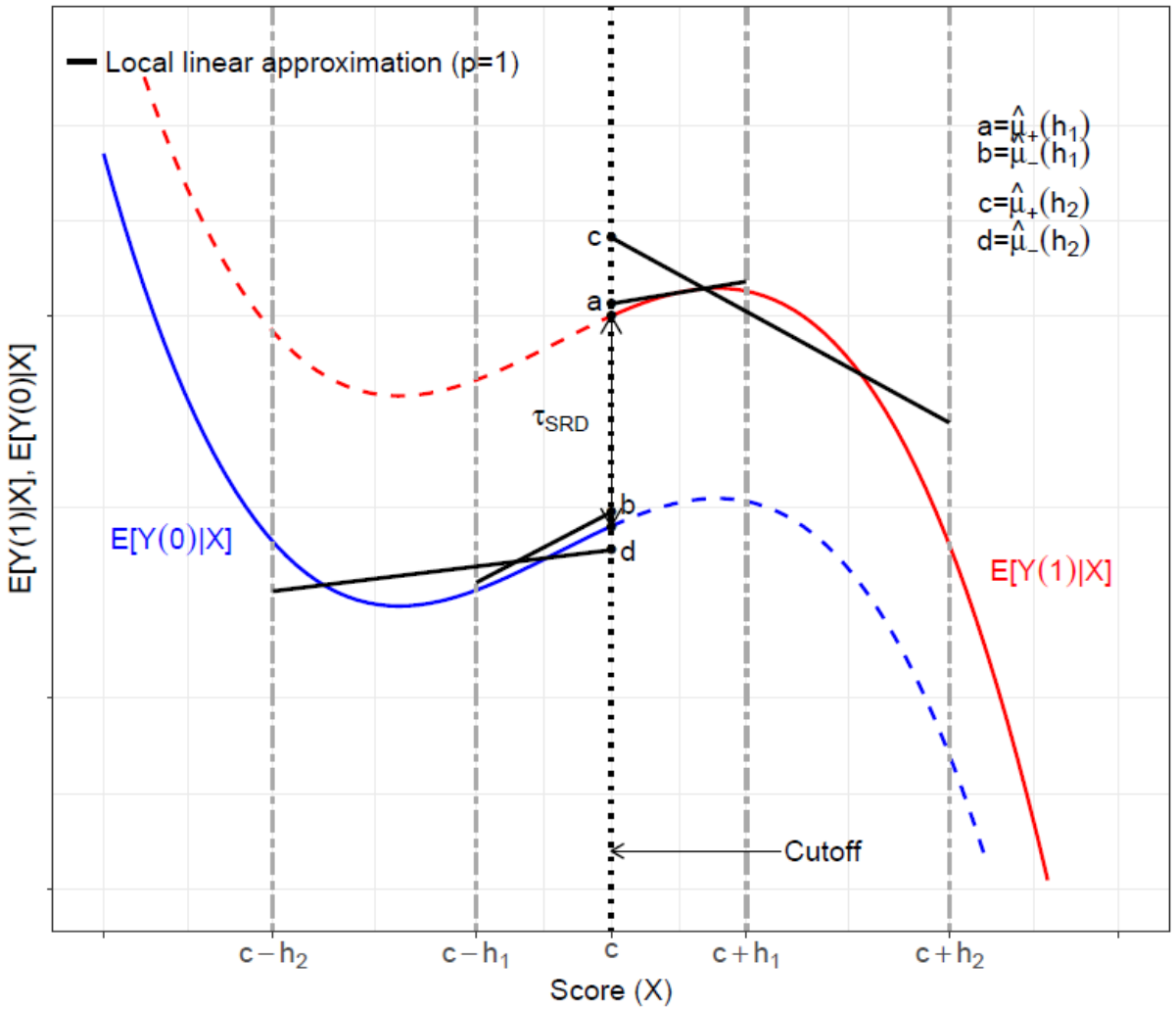
Important reminder: LOCAL effect

Figure 4.1: RD Estimation with local polynomial



Bias

Figure 4.3: Bias in Local Approximations



RDD: Polynomial Implementation Details

To make life easier:

- replace $\tilde{x}_i = x_i - c$.
- Estimate coefficients β: $(1, \tilde{x}, \tilde{x}^2, \ldots, \tilde{x}^p)$ and $\tilde{\beta}$: $(D_i, D_i\tilde{x}, D_i\tilde{x}^2, \ldots, D_i\tilde{x}^p)$.
- Now treatment effect at c just the coefficient on D_i. (We can ignore the interaction terms).
- If we want treatment effect at $x_i > c$ then we have to account for interactions.
  - Identification away from c is somewhat dubious.
- Lee and Lemieux (2010) suggest estimating a coefficient on a dummy for each bin in the polynomial regression $\sum_k \phi_k B_k$.
  - Add polynomials until you can satisfy the test that the joint hypothesis test that $\phi_1 = \cdots \phi_k = 0$.
  - There are better ways to choose polynomial order...

RDD: Checklist

Most RDD papers follow the same formula (so should yours)

- Plot of $P(D|X)$ so that we can see the discontinuity
- Plot of $E[Y|X]$ so that we see discontinuity there also
- Plot of $E[W|X]$ so that we don't see a discontinuity in controls.
- Density of X (check for manipulation).
- Show robustness to different "windows"
- The OLS RDD estimates
- The Local Linear RDD estimates
- The polynomial (from each side) RDD estimates
- An f-test of "bins" showing that the polynomial is flexible enough.

Read Lee and Lemieux (2010) before you get started.

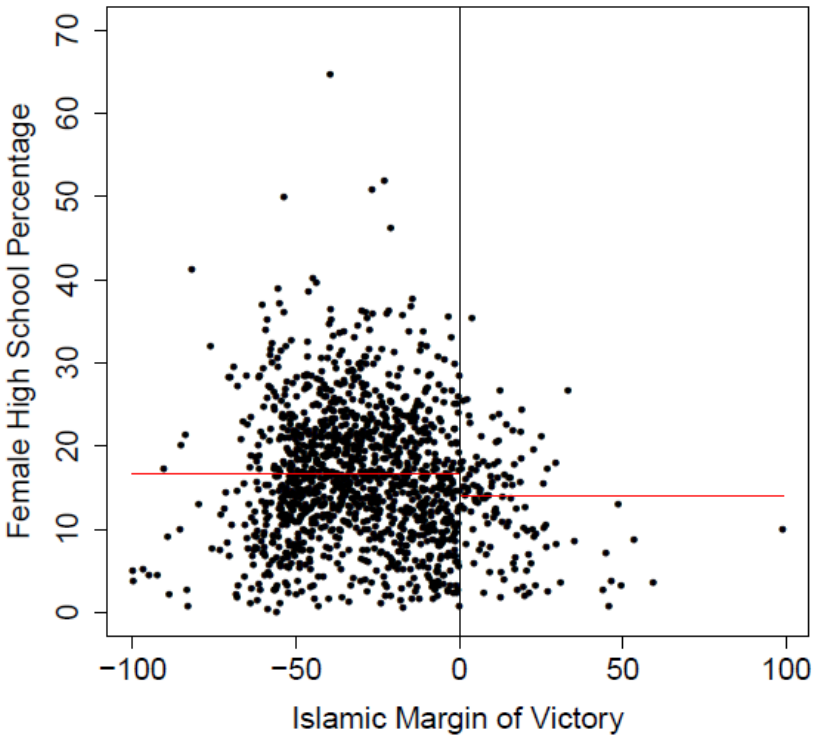
Application: Meyersson (ECMA, 2014)

- RQ: Does Islamic political control affect women's empowerment?
- Challenge: Islamic rule endogenous
- Meyerson uses the Lee instrument on 1994 Turkish minicipal elections
- Catteneo et al 2018 use this as a running example to demonstrate how to implement RD (and use their software)

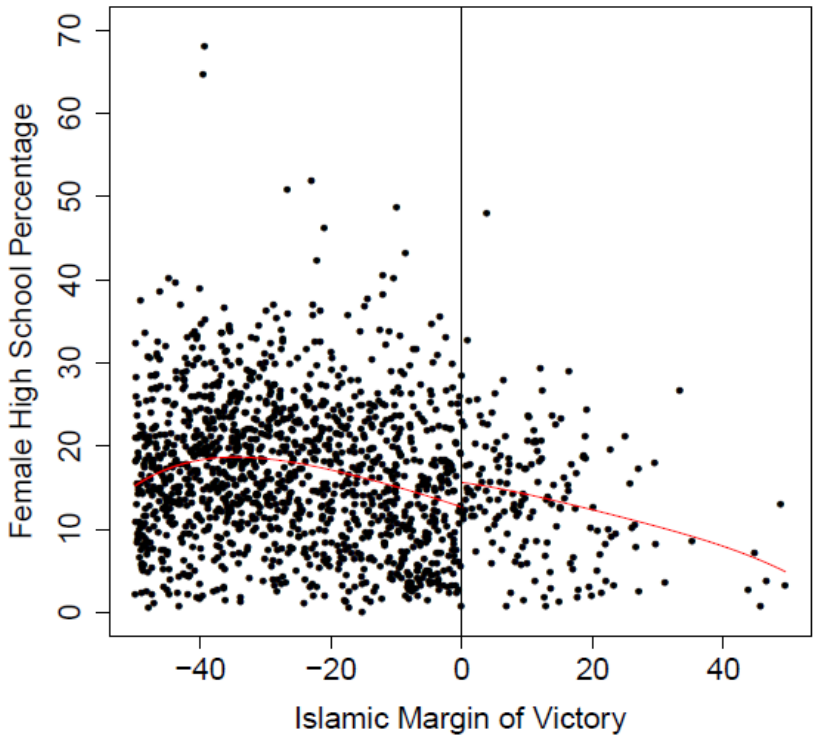
Note: For a similar replication exercise, check out the RDD chapter in Scott Cunningham's *The Mixtape*.

Raw vs Local Comparisons

Figure 2.3: Municipalities with Islamic Mayor vs. Municipalities with Secular Mayor—Meyersson data

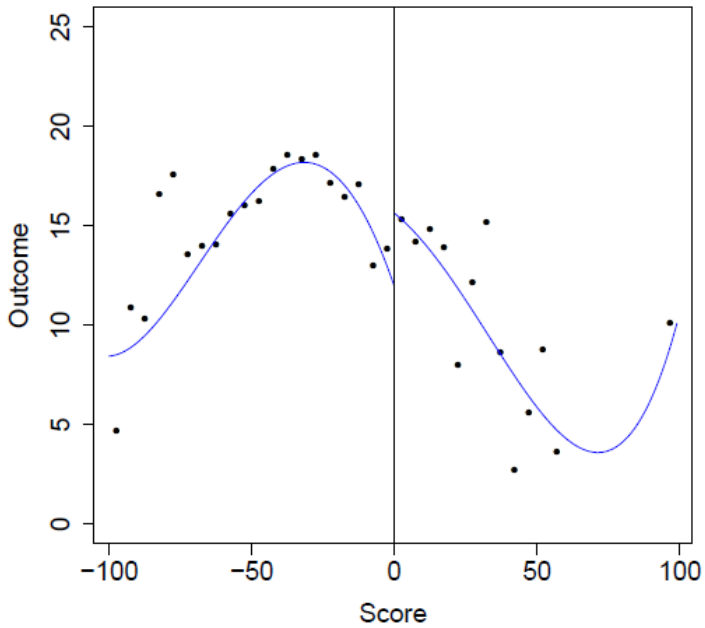


(a) Raw Comparison of Means

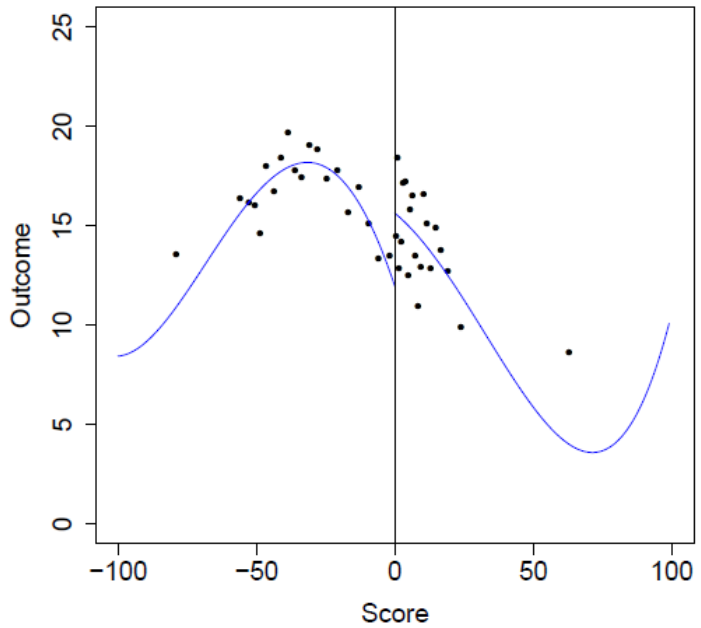


(b) Local Comparison of Means

Typically present bincatter



(a) 40 Evenly-Spaced Bins



(b) 40 Quantile-Spaced Bins

Challenges to identification

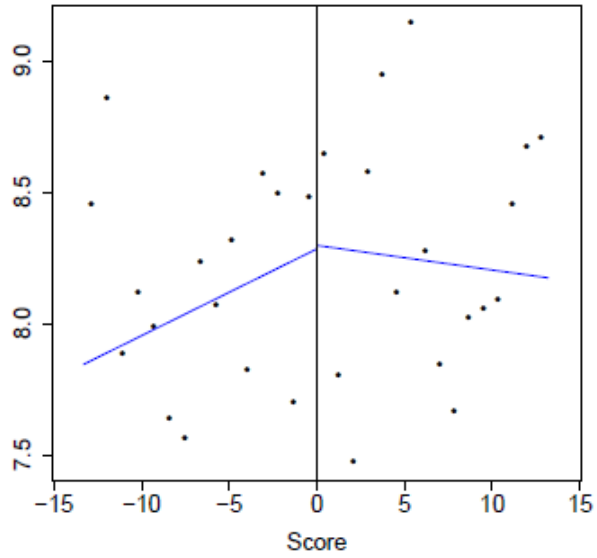
From The Mixtape

- The assignment rule is known in advance.
- Agents are interested in adjusting.
- Agents have time to adjust.
- The cutoff is endogenous to factors that independently cause potential outcomes to shift. (Lots of things change when people turn 18 / 21)
- There is nonrandom heaping along the running variable.

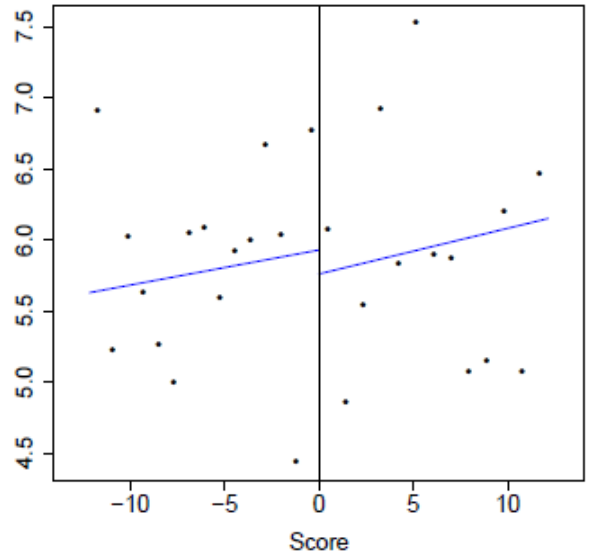
When to be worried? Individuals can "retake" the test, or self report the running variable.

Show other covariates smooth at cutoff

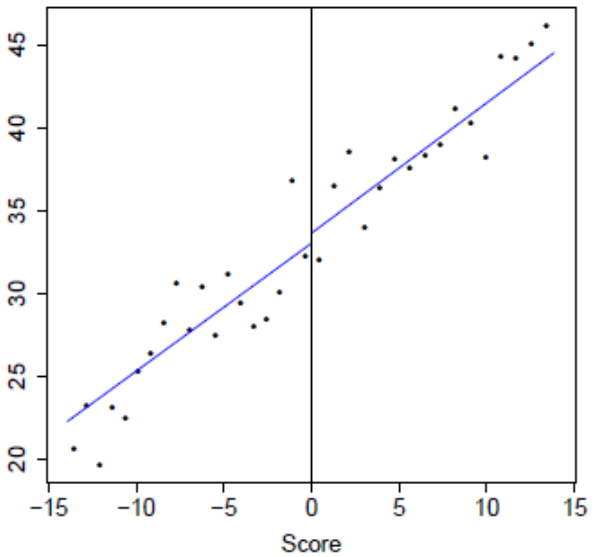
Figure 5.2: Graphical Illustration of Local Linear RD Effects for Predetermined Covariates—Meyersson data



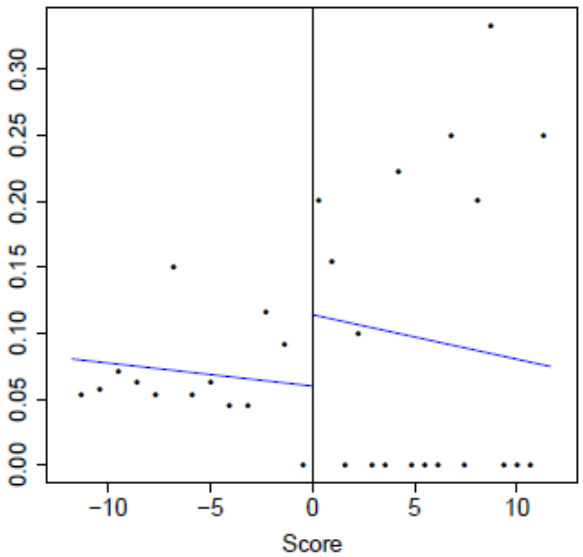
(a) Log Population in 1994



(b) Number of Parties Receiving Votes in 1994



(c) Islamic Vote Percentage in 1994



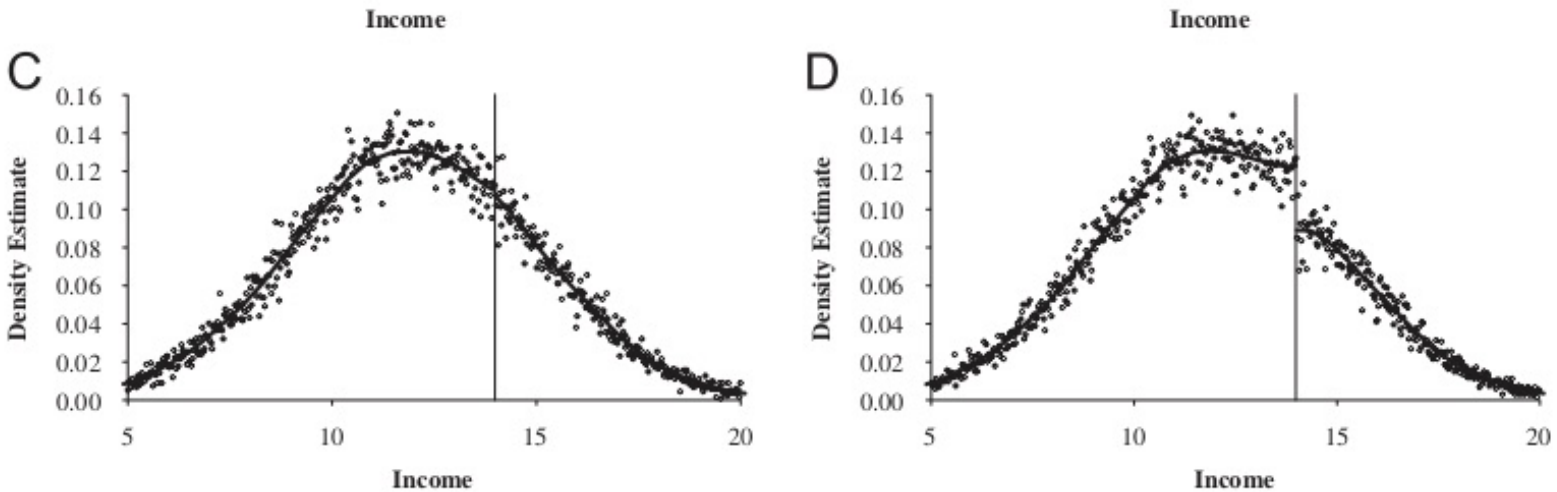
(d) Islamic Mayor in 1989

More generally, think of other placebo tests

- Shouldn't find jumps at the cutoff in other variables, or jumps in the same variable at other points (using the same estimator)

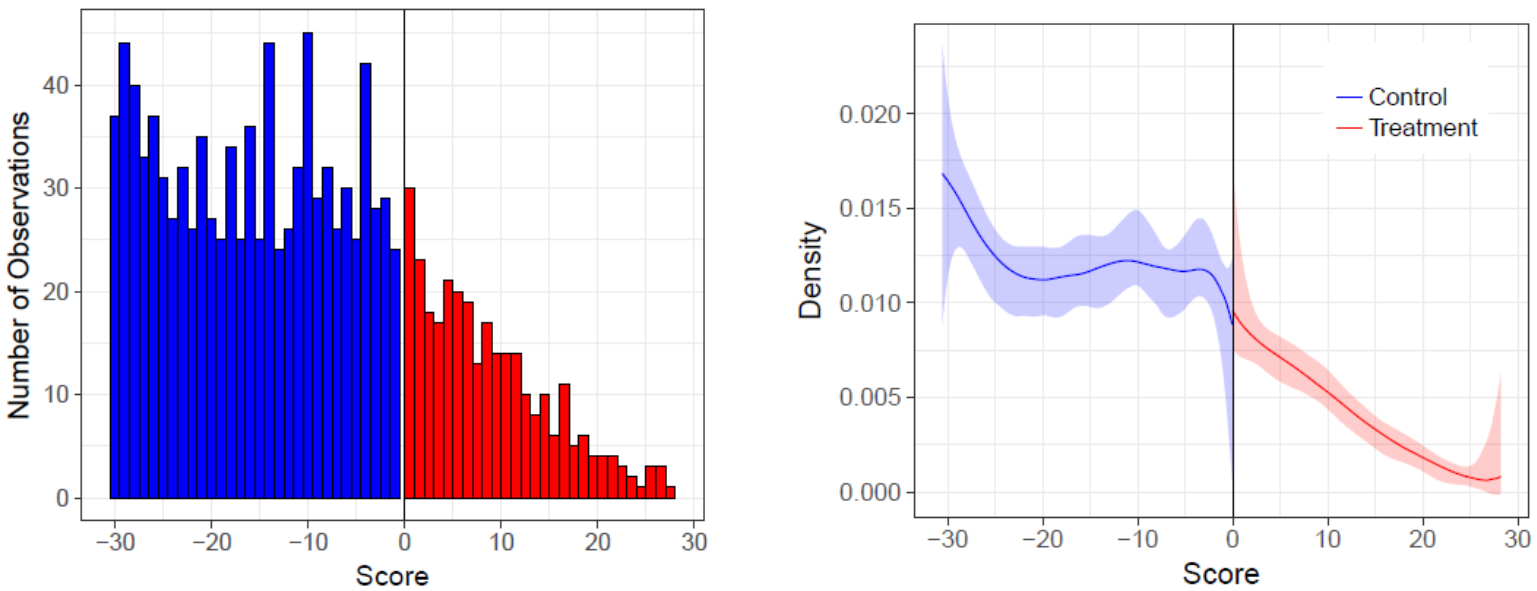
Look for bunching around the cutoff as evidence of manilability

- If agents are unable to manipulate the running variable, we should expect it to be smoothly distributed around the cutoff
- McCrary (2008) introduced a widely used test of continuity of the running variable density function



No evidence of bunching in the Islamic rule paper

Figure 5.4: Histogram and Estimated Density of the Score



(a) Histogram

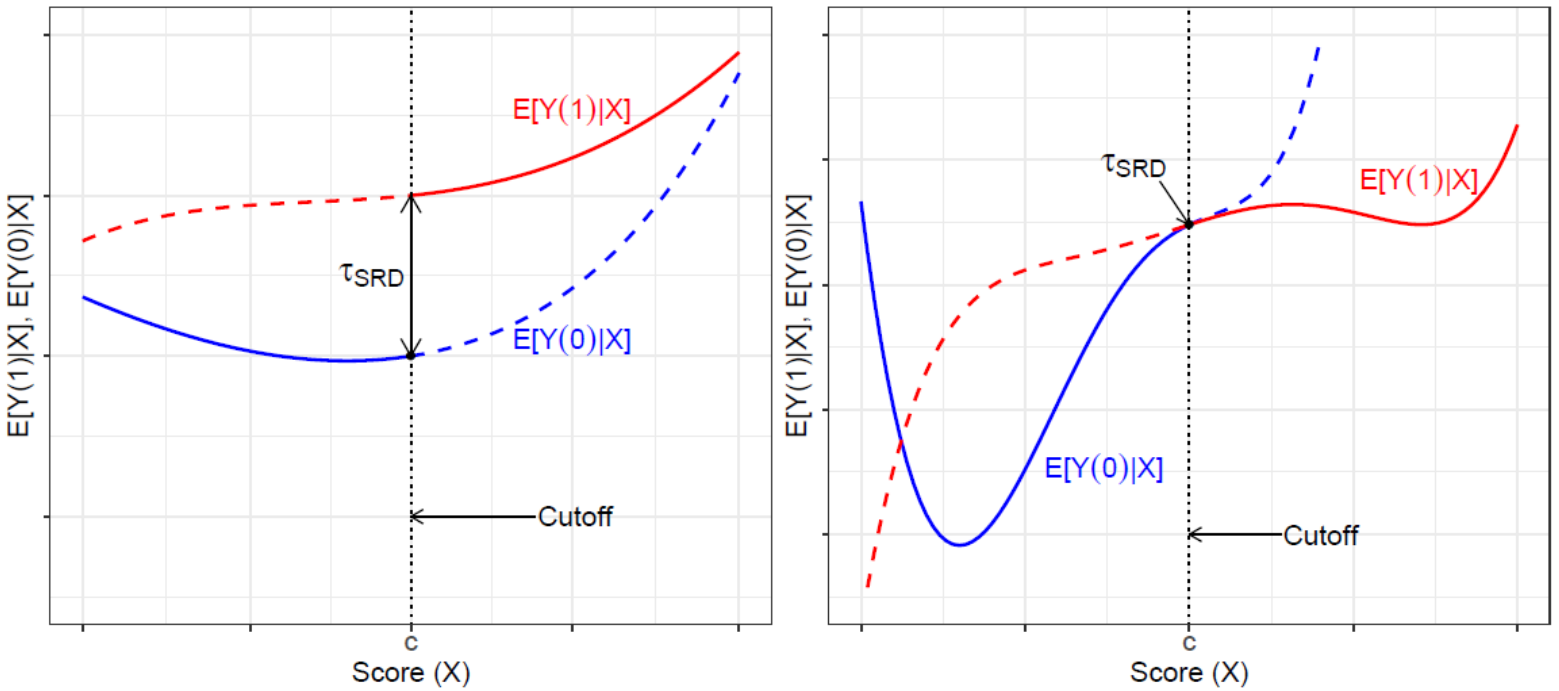
(b) Estimated Density

These tests and graphs can be easily implemented using the rdrobust package.

How useful is this LATE?

How useful is this LATE?

Figure 2.4: Local Nature of RD Effect



(a) Mild Heterogeneity

(b) Severe Heterogeneity

Other Examples

Luca on Yelp

- Have data on restaurant revenues and yelp ratings.
- Yelp produces a yelp score (weighted average rating) to two decimals ie: 4.32.
- Score gets rounded to nearest half star
- Compare 4.24 to 4.26 to see the impact of an extra half star.
- Now there are multiple discontinuities: Pool them? Estimate multiple effects?

Fuzzy RD

An important extension in the Fuzzy RD. Back to where we started:

$$\lim_{x \to c^+} P(T_i|X_i = x) \neq \lim_{x \to c^-} P(T_i|X_i = x)$$

- We need a discontinuous jump in probability of treatment, but it doesn't need to be $0 \to 1$.

$$\tau_i(c) = \frac{\lim_{x \to c^+} P(Y_i|X_i = x) - \lim_{x \to c^-} P(Y_i|X_i = x)}{\lim_{x \to c^+} P(T_i|X_i = x) - \lim_{x \to c^-} P(T_i|X_i = x)}$$

- Under sharp RD everyone was a complier, now we have some always takers and some never takers too.
- Now we are estimating the treatment effect only for the population of compliers at $x = c$.
- This should start to look familiar. We are going to do IV!

Related Idea: Kinks

A related idea is that of kinks.

- Instead of a discontinuous jump in the outcome there is a discontinuous jump in β_i on x_i.
- Often things like tax schedules or government benefits have a kinked pattern.

Back to treatment effect heterogeneity

- Consider a binary treatment T_i
- Potential outcomes

$$\begin{aligned} Y_{0i} &= \mu_0(X_i) + U_{0i} \\ Y_{1i} &= \mu_1(X_i) + U_{1i} \end{aligned}$$

- $\mu_j(x)$ represents the average outcome for individuals with observables x, and conditional mean zero U_{ji} captures (additively separable) unobserved heterogeneity
- Individual treatment effect $\tau_i = Y_{1i} - Y_{0i}$

Back to treatment effect heterogeneity

- Consider a binary treatment T_i
- Potential outcomes

$$\begin{aligned} Y_{0i} &= \mu_0(X_i) + U_{0i} \\ Y_{1i} &= \mu_1(X_i) + U_{1i} \end{aligned}$$

- $\mu_j(x)$ represents the average outcome for individuals with observables x, and conditional mean zero U_{ji} captures (additively separable) unobserved heterogeneity
- Individual treatment effect $\tau_i = Y_{1i} - Y_{0i}$
- ATE averages τ_i over entire population.
- ATT / ATU averages over those who received / didn't receive the treatment (somehow)
- LATE averages for those induced to switch due to an instrument

One quantity to rule them all: MTE

Heckman and Vytlacil (2005) provide a unifying non-parametric framework to categorize all of these: the marginal treatment effect or MTE

- Define the marginal treatment effect as the average treatment effect on the marginal individual entering treatment (ie not a number, this is a function).
- Key insight is that all of the other objects (LATE, ATE, ATT, etc.) can be written as integrals (weighted averages) of the MTE.
- The idea is to bridge the treatment effect parameters (stuff we get from running regressions) and the structural parameters: features of $f(\tau_i)$.

Setup: Selection into treatment

Consider the latent variable discrete choice problem

- Decision rule

$$T_i = 1(v_i \leq Z_i'\gamma)$$

  depends on at least one instrument Z which does not affect potential outcomes.
- V_i represents the unobserved disutility of decision.
- For a given $Z_i'\gamma$ and individual is marignal if $Z_i'\gamma = v_i$
- Consider the TE $\tau_i = Y_i^1 - Y_i^0$
- The MTE is $MTE(Z_i) = E(\tau_i | v_i = Z_i'\gamma)$

Example: Selection into college

Let τ_i be lifetime earnings with and without college.

- Costs of attending $C_i = w_0 + \gamma' z_i + v_i$
- Rational students attend if

$$\tau_i - [w_0 + \gamma' z_i + v_i] > 0$$

- If we could condition on marginal students,

$$v_i = \tau_i - [w_0 + \gamma' z_i]$$

we'd be able to pin down the treatment effect at a given Z_i

Propensity score

- Individuals select into treament if the observable portion of their decision utility exceeds their unobserved resistance V_i
- Let F_V be the cdf of this resistance. Observed treatment thus implies

$$F_V(\mu_T(X_i, Z_i)) \geq F_V(V_i)$$

- As written, the LHS is just the propensity score: the probability of treatment based on observables.
- The RHS is simply individual i's quantile of the unobserved distaste distribution. Let $u_{si} = F_V(V_i)$.
- So if an indivdual with a propensity score $P(X_i, Z_i) = p$ selects into treatment, it must be that that individual V_i is in the bottom pth percentile of the V distribution.

MTE: Derivation

For simplicity write

$$\begin{aligned} Y_{0i} &= \gamma_0' X_i + U_{0i} \\ Y_{1i} &= \gamma_1' X_i + U_{1i} \end{aligned}$$

For any individual we observe

$$Y_i = \gamma_0' X_i + T_i(\gamma_1 - \gamma_0)' X_i + U_{0i} + T_i(U_{1i} - U_{0i})$$

Think of the propensity score as an instrument

$P(T = 1|Z) = P(Z)$ works as our instrument with two assumptions:

1. $(U_0, U_1, u_s) \perp P(Z)|X$. (Exogeneity)
2. $P(Z|X)$ continuous support – ie conditional on X there is enough variation in Z for $P(Z)$ to take on all values $\in (0, 1)$.
   - This is much stronger than typical relevance condition.

MTE: Derivation

Take the expectation conditional on x and the instrument

$$E[Y|X, P(Z) = p] = \gamma_0' X + p(\gamma_1 - \gamma_0)' X + E[T(U_1 - U_0)|X, P(Z) = p]$$

MTE: Derivation

Take the expectation conditional on x and the instrument

$$E[Y|X, P(Z) = p] = \gamma_0' X + p(\gamma_1 - \gamma_0)' X + E[T(U_1 - U_0)|X, P(Z) = p]$$

Note that $T = 1$ over the interval $u_s = [0, p]$ and zero for higher values of u_s.

$$E[T(U_1 - U_0)|P(Z) = p, X] = \int_{-\infty}^{\infty} \int_{0}^{p} (U_1 - U_0) f((U_1 - U_0)|U_s = u_s) du_s d(U_1 - U_0)$$

MTE: Derivation

Let $U_1 - U_0 \equiv \eta$. Can now express the MTE as

$$\begin{aligned}\Delta^{MTE}(p) &= \frac{\partial E[Y|X, P(Z) = p]}{\partial p} \\ &= (\gamma_1 - \gamma_0)'X + \int_{-\infty}^{\infty} \eta f(\eta|U_s = p) d\eta \\ &= (\gamma_1 - \gamma_0)'X + E[\eta|u_s = p]\end{aligned}$$

What is $E[\eta|u_s = p]$? The expected unobserved gain from treatment of those people whose unobserved characteristics make them indifferent between treatment at $P(Z) = p$ (ie who are at the margin).

How to Estimate an MTE

1. Estimate $P(Z) = Pr(T = 1|Z)$ nonparametrically (include exogenous part of X in Z).
2. Nonparametric regression of Y on X and $P(Z)$
3. For example,

$$E[Y|X, P(Z)p] = \gamma_0' X + \hat{p}(\gamma_1 - \gamma_0)' X + \kappa(\hat{p})$$

   where $\kappa()$ is some nonlinear function (polynomials?)
4. Differentiate w.r.t. $P(Z)$
5. plot it for all values of $P(Z) = p$.

So long as $P(Z)$ covers $(0, 1)$ then we can trace out the full distribution of $\Delta^{MTE}(p)$.

Can now define any average we want in terms of MTE

Calculate the outcome given (X, Z) (actually X and $P(Z) = p$).
ATE : This one is obvious. We treat everyone!

$$\int_{-\infty}^{\infty} \Delta^{MTE}(p) = (\gamma_1 - \gamma_0)'X + \underbrace{\int_{-\infty}^{\infty} E(\eta|u_s) d\, u_s}_{0}$$

What about LATE?

- LATE: Fix an X and $P(Z)$
- Consider a policy which varies probability of treatment for X from $b(X)$ to $a(X)$ with $a > b$.
- LATE integrates over the compliers with $b(X) \leq u_s \leq a(X)$.

$$\begin{aligned} LATE(X) &= \int_{-\infty}^{\infty} \Delta^{MTE}(p) \\ &= (\gamma_1 - \gamma_0)'X + \frac{1}{a(X) - b(X)} \int_{b(X)}^{a(X)} E(\eta | u_s) d\, u_s \end{aligned}$$

- One thing to note is that obviously LATE depends on the margin the policy shifts

How does this compare to IV?

[Some of what follows comes from Cornelissen et al. (2016)]

- Consider the Wald estimator with two points from a continous instrument

$$Wald(z, z', x) = \frac{E[Y_i|Z_i = z, X_i = x] - E[Y_i|Z_i = z', X_i = x]}{E[T_i|Z_i = z, X_i = x] - E[Y_i|Z_i = z', X_i = x]}$$

- We showed this recovers

$$\begin{aligned} LATE(z, z', x) &= E[\tau_i|T_{iz} > T_{iz'}, X_i = x] \\ &= E[\tau_i|P(z') < u_s < P(z), X_i = x] \end{aligned}$$

Consider a discretization of treatment probability distribution

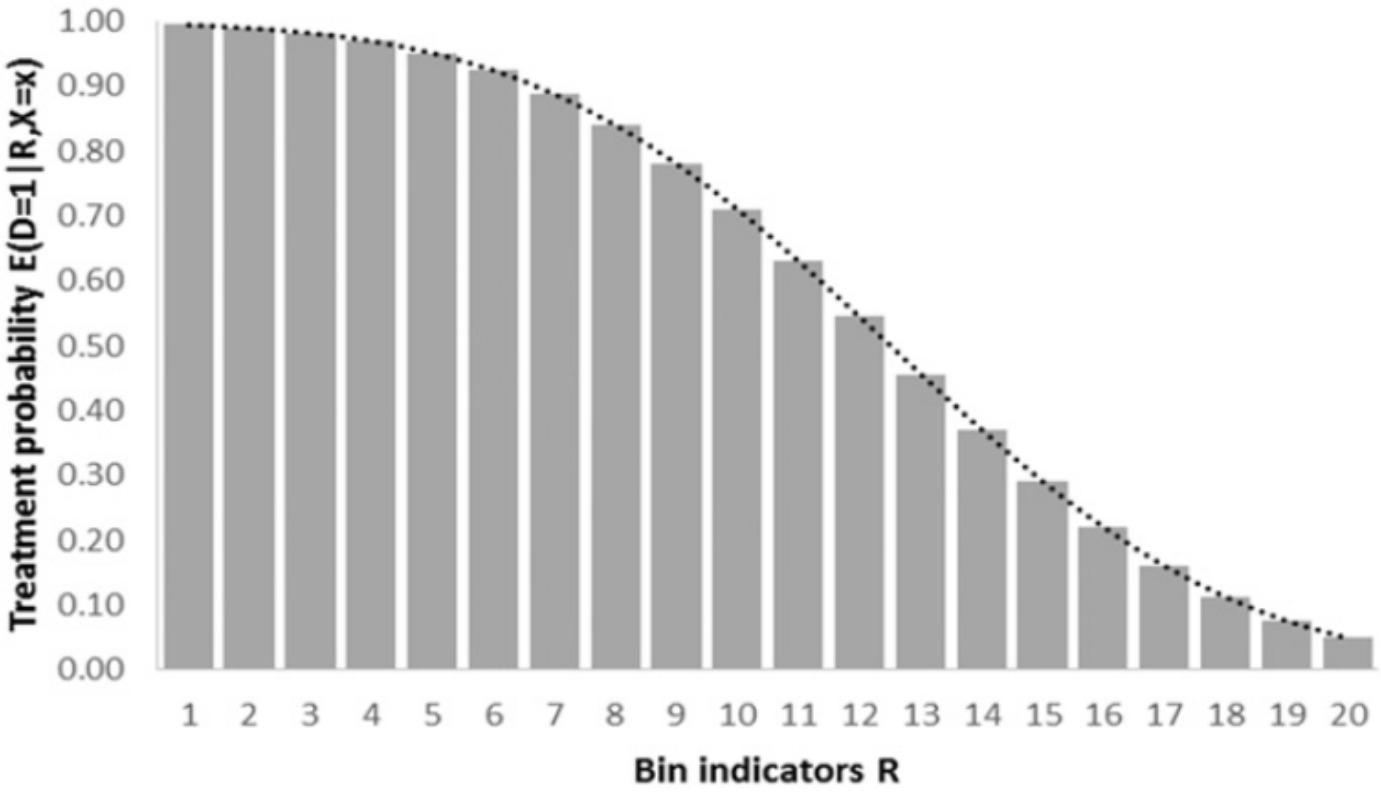


Fig. 2. Treatment probability in discrete bins of a continuous instrument. *Notes:* Based on hypothetical data, the bins in this figure show the probability of treatment in a sample with fixed covariates ($E[D = 1,R,X = x]$) as a function of a discrete variable R, which has been generated by grouping the values of the continuous instrument depicted in Fig. 1 into 20 equally spaced bins. The dotted line reproduces the function depicted in Fig. 1. *Data source:* Simulated hypothetical data.

Grouped IV averages relationship across bins

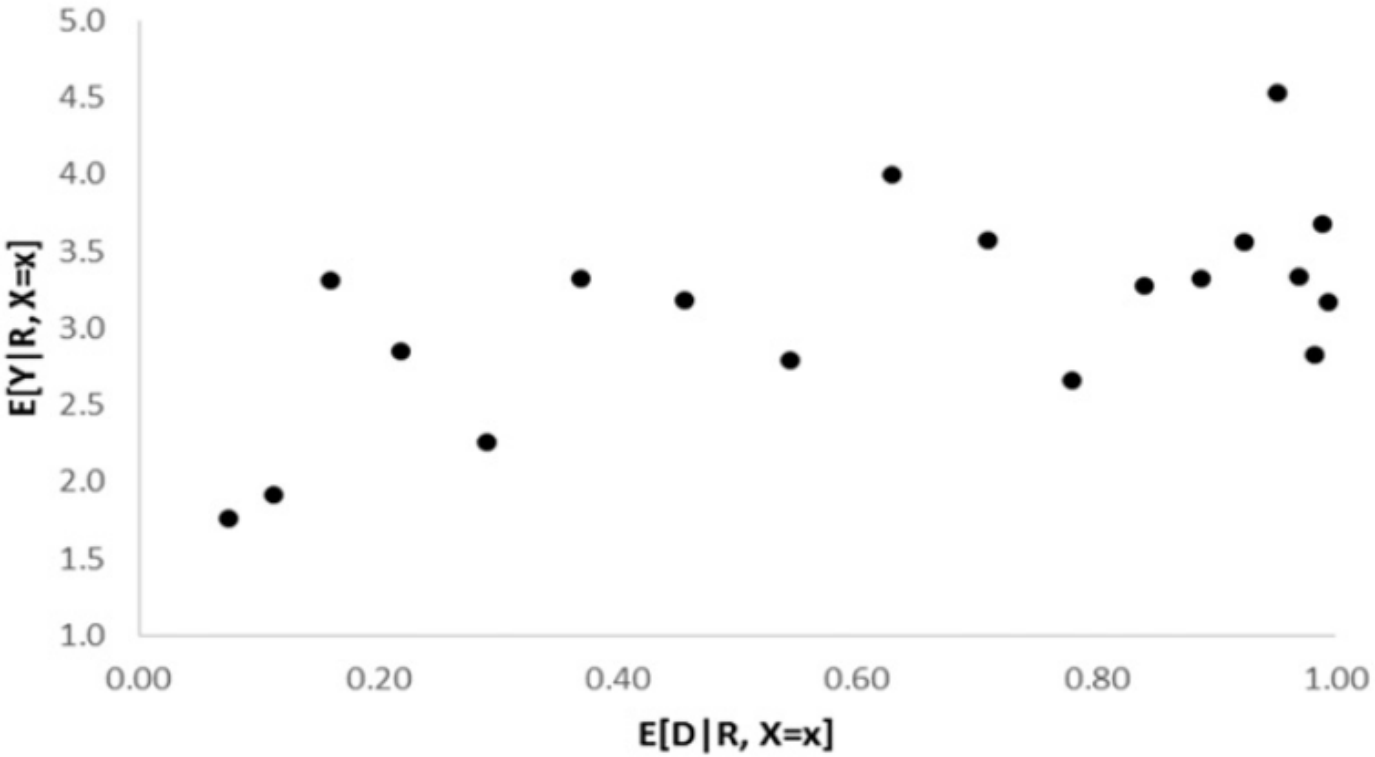


Fig. 3. Grouped data IV. *Notes:* Based on hypothetical data, the figure plots the average outcome against the average treatment probability in a sample with fixed covariates for 20 groups, which are equal to the bins depicted in Fig. 2 and correspond to 20 equally ument. Grouped data IV can be visualized as *e:* Simulated hypothetical data.

How does this compare to IV?

- 2SLS is going to fit a line through these heterogenous effects, and aggregate IV will be the slope.
- MTE allows slope to vary across very fine bins
- Looking at the Wald formula, can see that MTE for a given u_s is the limit of LATE as $P(z') \to P(z)$
- Thus the MTE is actually identified from local IV (LIV) using small departures from propensity score at $u_s = P(z)$

How does this compare to IV?

- 2SLS is going to fit a line through these heterogenous effects, and aggregate IV will be the slope.
- MTE allows slope to vary across very fine bins
- Looking at the Wald formula, can see that MTE for a given u_s is the limit of LATE as $P(z') \to P(z)$
- Thus the MTE is actually identified from local IV (LIV) using small departures from propensity score at $u_s = P(z)$

HV ECMA 2005 show everything is a weighted MTE

Table 2A

Treatment effects and estimands as weighted averages of the marginal treatment effect

---

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

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

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

Policy relevant treatment effect: $\text{PRTE}(x) = E(Y_{a'} \mid X = x) - E(Y_a \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\omega_{\text{PRTE}}(x, u_D)\, du_D$ for two policies a and a' that affect the Z but not the X

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

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

Source: Heckman and Vytlacil (2005).

HV ECMA 2005 show everything is a weighted MTE

Table 2B
Weights

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

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

$$\omega_{\text{TUT}}(x, u_D) = \left[\int_{0}^{u_D} f_{P|X}(p \mid X = x)\, dp\right] \frac{1}{E((1-P)|X=x)}$$

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

$$\Delta\overline{P}(x) = E(P_a \mid X = x) - E(P_{a'} \mid X = x)$$

$$\omega^{J}_{\text{IV}}(x, u_D) = \left[\int_{u_D}^{1} (J(Z) - E(J(Z) \mid X = x)) f_{J,P|X}(j, t \mid X = x)\, dt\, dj\right] \frac{1}{\text{Cov}(J(Z), D|X=x)}$$

$$\omega_{\text{OLS}}(x, u_D) = 1 + \frac{E(U_1|X=x, U_D=u_D)\omega_1(x,u_D) - E(U_0|X=x, U_D=u_D)\omega_0(x,u_D)}{\Delta^{\text{MTE}}(x,u_D)}$$

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

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

HV Example: Roy Model

The model

Outcomes	Choice model
$Y_1 = \mu_1 + U_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \mu_0 + U_0 = \alpha + U_0$	$D = \begin{cases} 1 & \text{if } D^* \geqslant 0, \\ 0 & \text{if } D^* < 0 \end{cases}$

General case

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

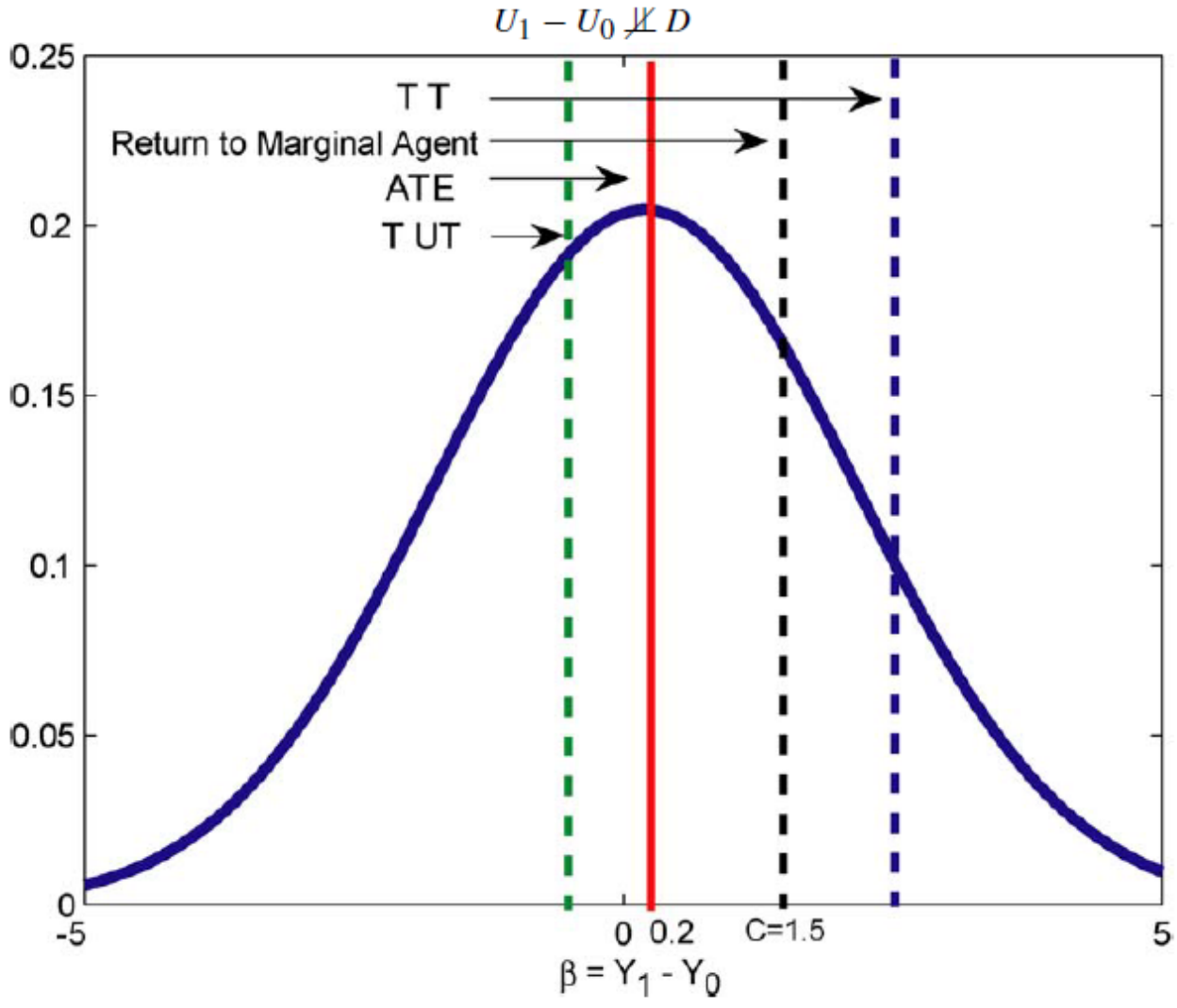
The researcher observes (Y, D, C).
$Y = \alpha + \beta D + U_0$ where $\beta = Y_1 - Y_0$.

Parameterization

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

Figure 1. Distribution of gains in the Roy economy. *Source*: Heckman, Urzua and Vytlacil (2006).

HV Roy Model Averages

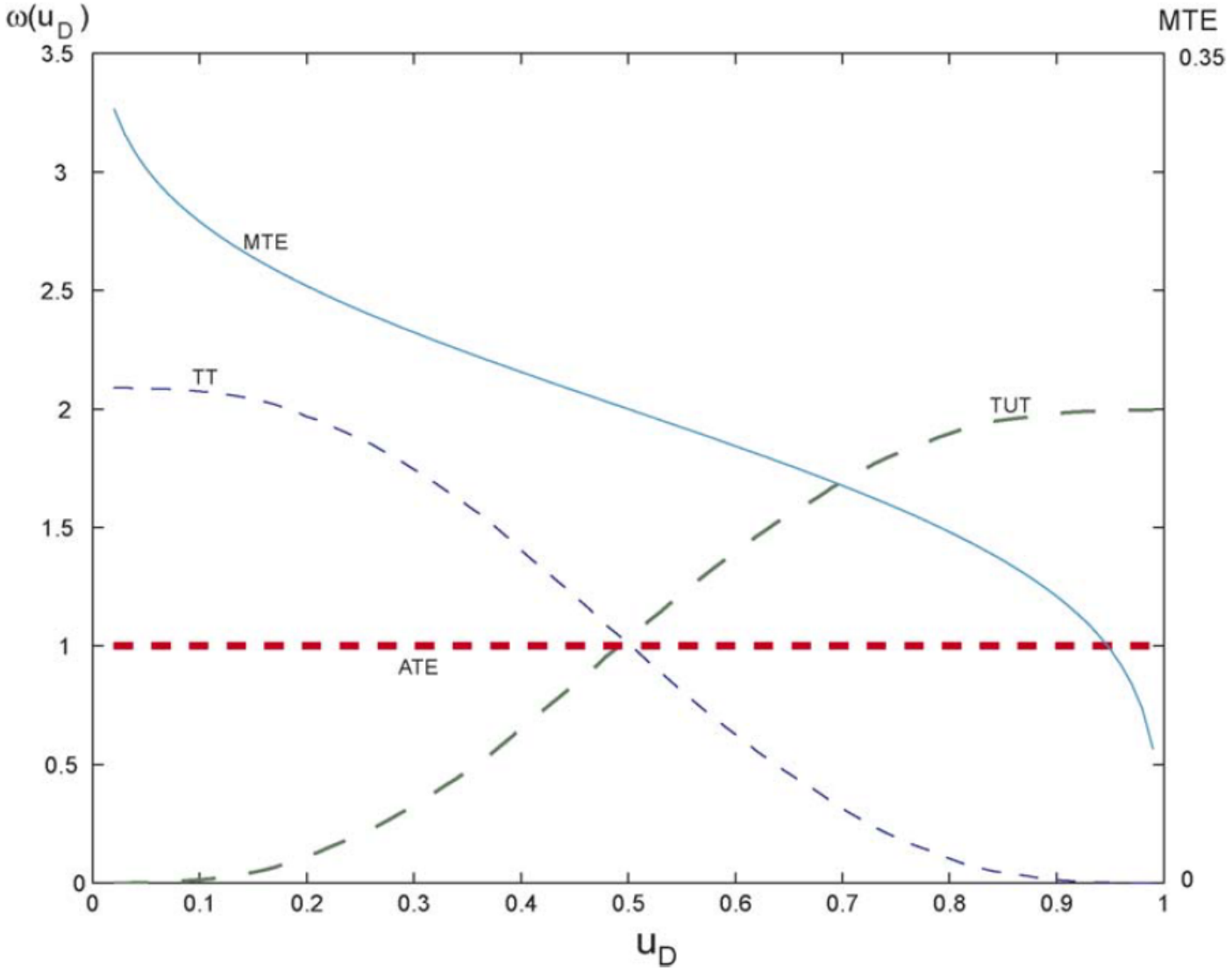


$TT = 2.666$, $TUT = -0.632$

Return to marginal agent $= C = 1.5$

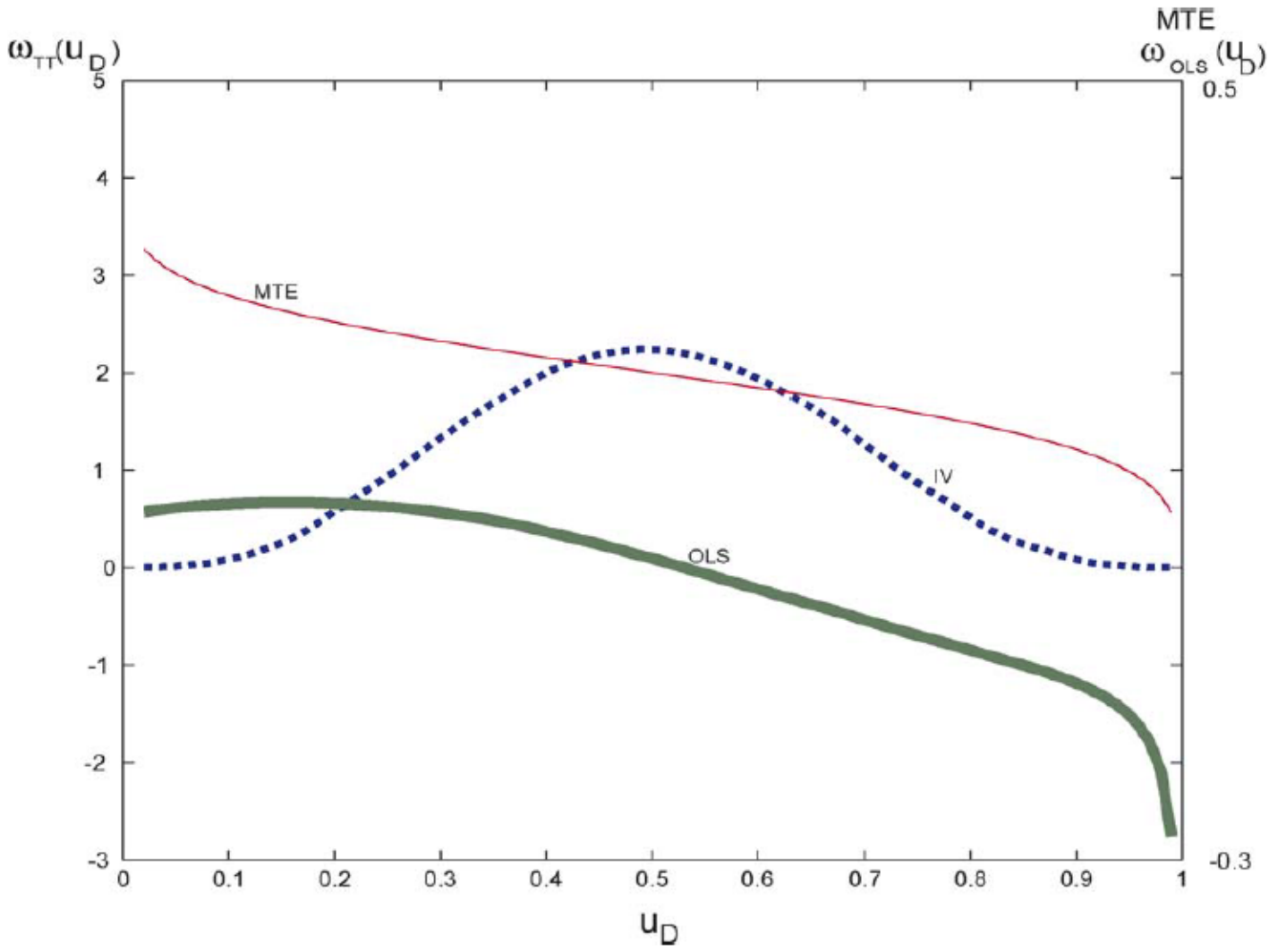
$ATE = \mu_1 - \mu_0 = \bar{\beta} = 0.2$

HV Roy Model: Weights



IV vs OLS weights

Under monotonicity IV weights all positive.
No guarantee for OLS.



$$Y_1 = \alpha + \bar{\beta} + U_1 \qquad U_1 = \sigma_1 \tau \qquad \alpha = 0.67 \qquad \sigma_1 = 0.012$$
$$Y_0 = \alpha + U_0 \qquad U_0 = \sigma_0 \tau \qquad \bar{\beta} = 0.2 \qquad \sigma_0 = -0.050$$

Carneiro, Heckman, and Vytlacil (2011)

- Estimate returns to college (including heterogeneity of returns). $T_i = 1$ if ever attended college.
- NLSY 1979
- $Y = \log(wage)$ in 1991
- Covariates X: Experience (years), Ability (AFQT Score), Mother's Education, Cohort Dummies, State Unemployment, MSA level average wage.
- Instruments Z:
  - Cost shifters: College in MSA ; In state cost
  - Opportunity cost: average earnings in MSA and avg unemployment (at 17).

Propensity estimate: Logit

TABLE 3—COLLEGE DECISION MODEL: AVERAGE MARGINAL DERIVATIVES

	Average derivative
Controls (**X**)	
Corrected AFQT	0.2826
	(0.0114)***
Mother's years of schooling	0.0441
	(0.0059)***
Number of siblings	−0.0233
	(0.0068)***
Urban residence at 14	0.0340
	(0.0274)
"Permanent" local log earnings at 17	0.1820
	(0.0941)**
"Permanent" state unemployment rate at 17	0.0058
	(0.0165)
Instruments (**Z**)	
Presence of a college at 14	0.0529
	(0.0273)**
Local log earnings at 17	−0.2687
	(0.1008)***
Local unemployment rate at 17 (in percent)	0.0149
	(0.0100)
Tuition in 4 year public colleges at 17 (in $100)	−0.0027
	(0.0017)*
Test for joint significance of instruments: *p*-value	0.0001

Notes: This table reports the coefficients and average marginal derivatives from a logit regres-

Carneiro, Heckman and Vytlacil

TABLE 4— TEST OF LINEARITY OF $E(Y|\mathbf{X}, P = p)$ USING POLYNOMIALS IN P; AND TEST OF EQUALITY OF LATEs OVER DIFFERENT INTERVALS (H_0: $LATE^j\,(U_S^{Lj}, U_S^{Hj}) - LATE^{j+1}\,(U_S^{Lj+1}, U_S^{Hj+1}) = 0$)

Panel A. Test of linearity of $E(Y|\mathbf{X}, P = p)$ using models with different orders of polynomials in P[a]

Degree of polynomial for model	2	3	4	5
p-value of joint test of nonlinear terms	0.035	0.049	0.086	0.122
Adjusted critical value		0.057		
Outcome of test		Reject		

Panel B. Test of equality of LATEs (H_0: $LATE^j\,(U_S^{Lj}, U_S^{Hj}) - LATE^{j+1}\,(U_S^{Lj+1}, U_S^{Hj+1}) = 0$)[b]

Ranges of U_S for $LATE^j$	(0, 0.04)	(0.08, 0.12)	(0.16, 0.20)	(0.24, 0.28)	(0.32, 0.36)	(0.40, 0.44)
Ranges of U_S for $LATE^{j+1}$	(0.08, 0.12)	(0.16, 0.20)	(0.24, 0.28)	(0.32, 0.36)	(0.40, 0.44)	(0.48, 0.52)
Difference in LATEs	0.0689	0.0629	0.0577	0.0531	0.0492	0.0459
p-value	0.0240	0.0280	0.0280	0.0320	0.0320	0.0520
Ranges of U_S for $LATE^j$	(0.48, 0.52)	(0.56, 0.60)	(0.64, 0.68)	(0.72, 0.76)	(0.80, 0.84)	(0.88, 0.92)
Ranges of U_S for $LATE^{j+1}$	(0.56, 0.60)	(0.64, 0.68)	(0.72, 0.76)	(0.80, 0.84)	(0.88, 0.92)	(0.96, 1)
Difference in LATEs	0.0431	0.0408	0.0385	0.0364	0.0339	0.0311
p-value	0.0520	0.0760	0.0960	0.1320	0.1800	0.2400
Joint p-value			0.0520			

CHV Normal Selection Model

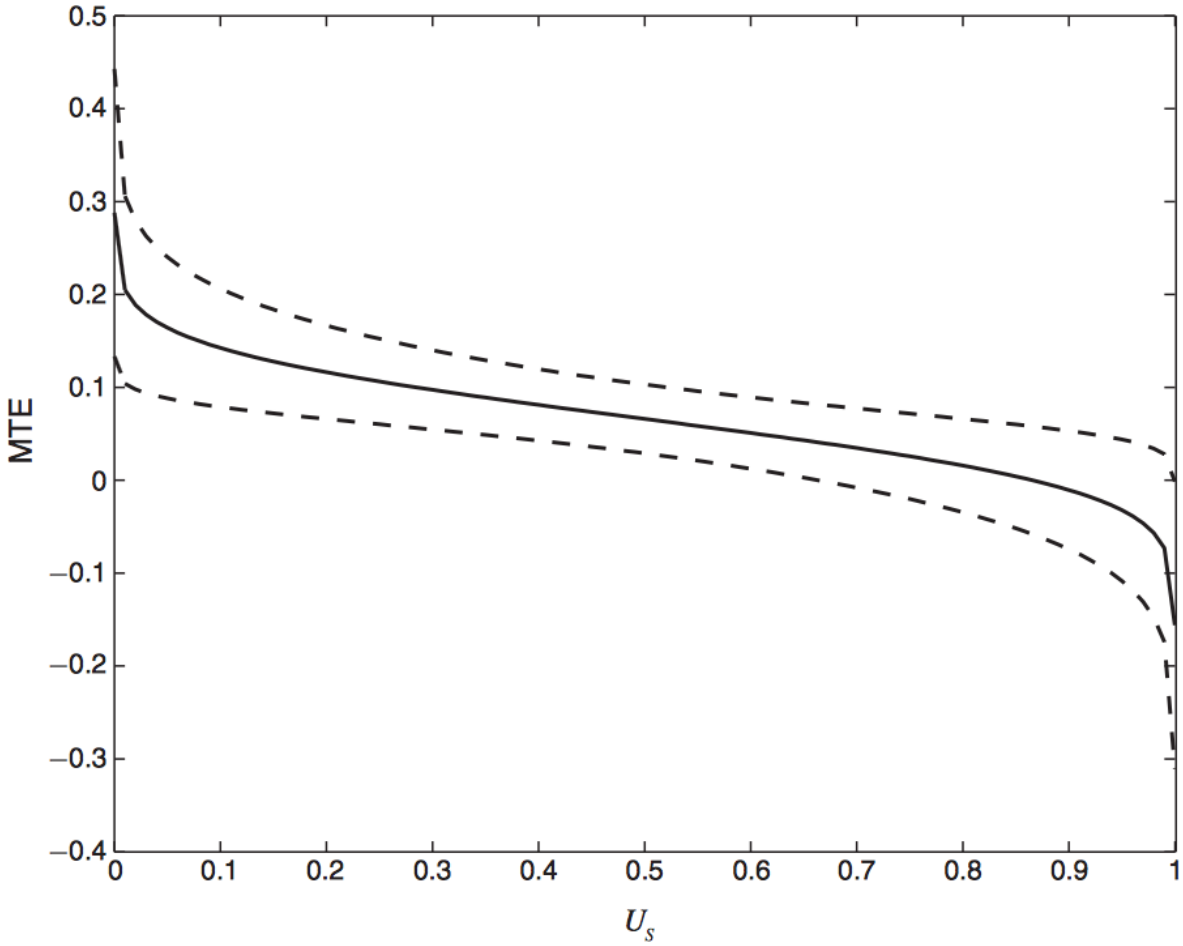


FIGURE 1. MTE ESTIMATED FROM A NORMAL SELECTION MODEL

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:

$$\Delta^{\text{MTE}}(\mathbf{x}, u_S) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) - (\sigma_{1V} - \sigma_{0V})\,\Phi^{-1}(u_S),$$

where σ_{1V} and σ_{0V} are the covariances between the unobservables of the college and high school equation and the

Carneiro, Heckman and Vytlacil

TABLE 5—RETURNS TO A YEAR OF COLLEGE

Model		Normal	Semiparametric
$ATE = E(\beta)$		0.0670	Not identified
		(0.0378)	
$TT = E(\beta \mid S = 1)$		0.1433	Not identified
		(0.0346)	
$TUT = E(\beta \mid S = 0)$		−0.0066	Not identified
		(0.0707)	
	MPRTE		
Policy perturbation	Metric		
$Z^k_\alpha = Z^k + \alpha$	$\lvert \mathbf{Z}\gamma - V \rvert < e$	0.0662	0.0802
		(0.0373)	(0.0424)
$P_\alpha = P + \alpha$	$\lvert P - U \rvert < e$	0.0637	0.0865
		(0.0379)	(0.0455)
$P_\alpha = (1 + \alpha)P$	$\lvert \frac{P}{U} - 1 \rvert < e$	0.0363	0.0148
		(0.0569)	(0.0589)
Linear IV (Using $P(\mathbf{Z})$ as the instrument)		0.0951	
		(0.0386)	
OLS		0.0836	
		(0.0068)	

Notes: This table presents estimates of various returns to college, for the semiparametric and the normal selection models: average treatment effect (ATE), treatment on the treated (TT), treatment on the untreated (TUT), and different versions of the marginal policy relevant treatment effect (MPRTE). The linear IV estimate uses P as the instrument. Standard errors are

CHV Local IV MTE

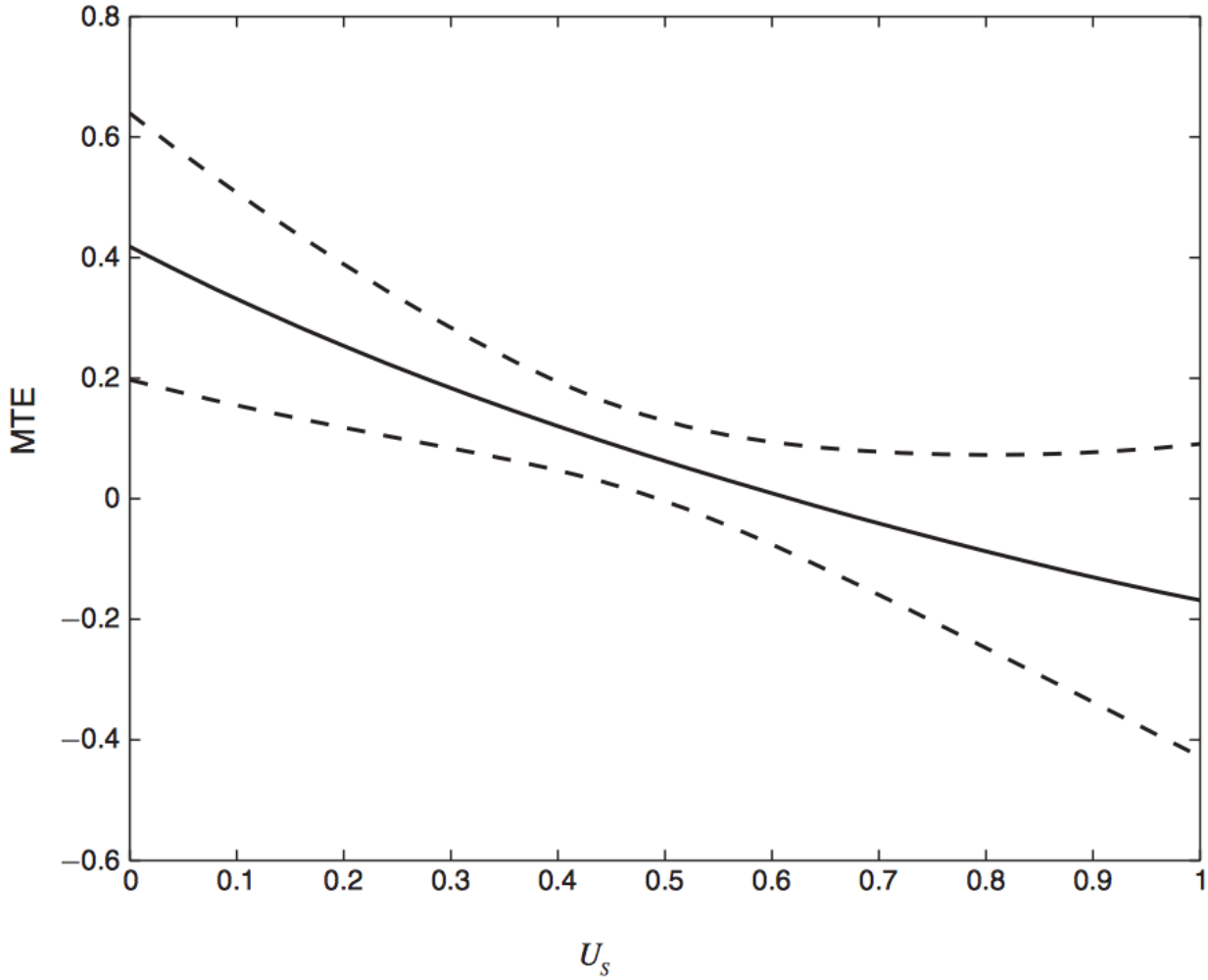


FIGURE 4. $E(Y_1 - Y_0 | \mathbf{X}, U_S)$ WITH 90 PERCENT CONFIDENCE INTERVAL—LOCALLY QUADRATIC REGRESSION ESTIMATES

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

CHV Local IV MTE

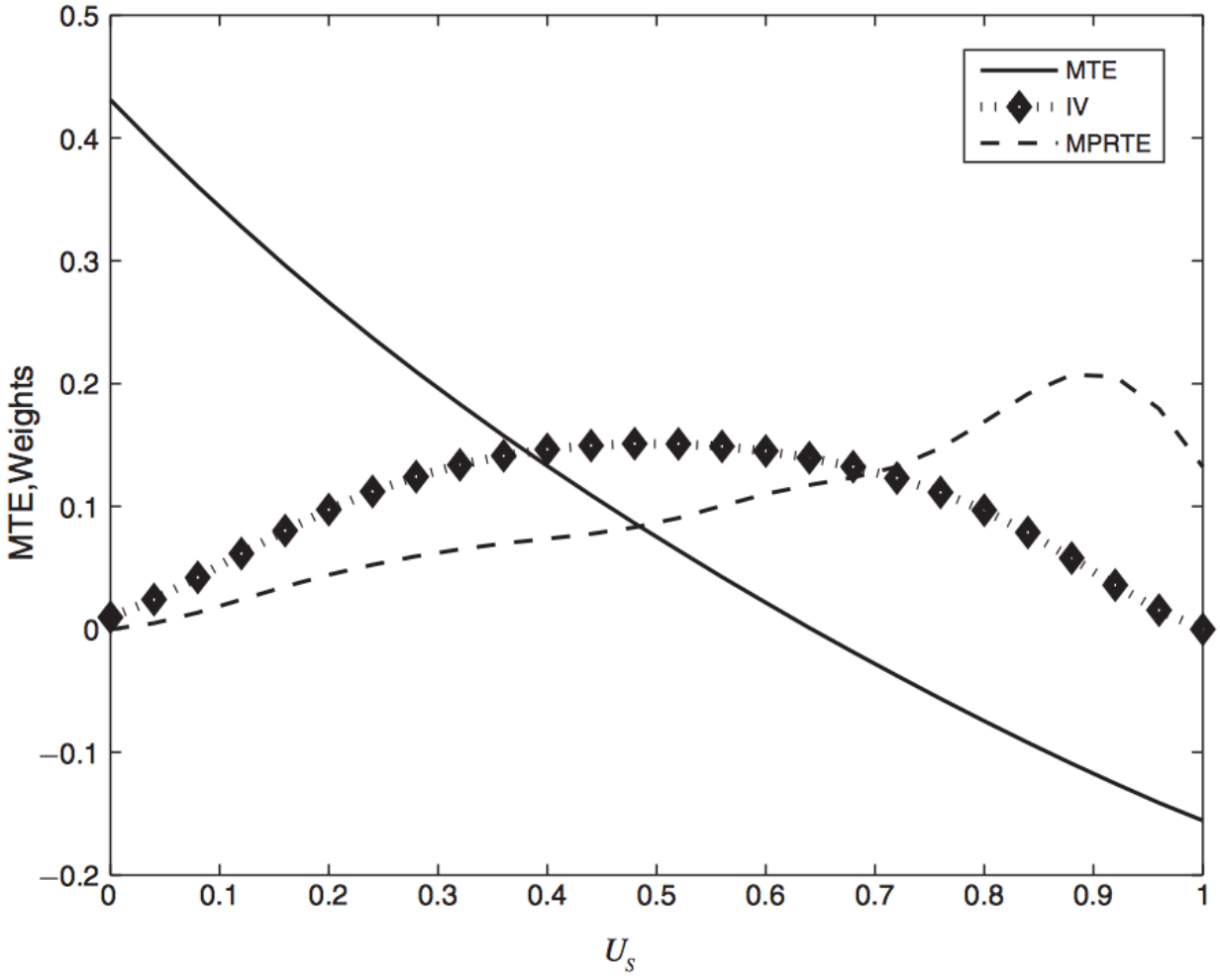


FIGURE 6. WEIGHTS FOR IV AND MPRTE

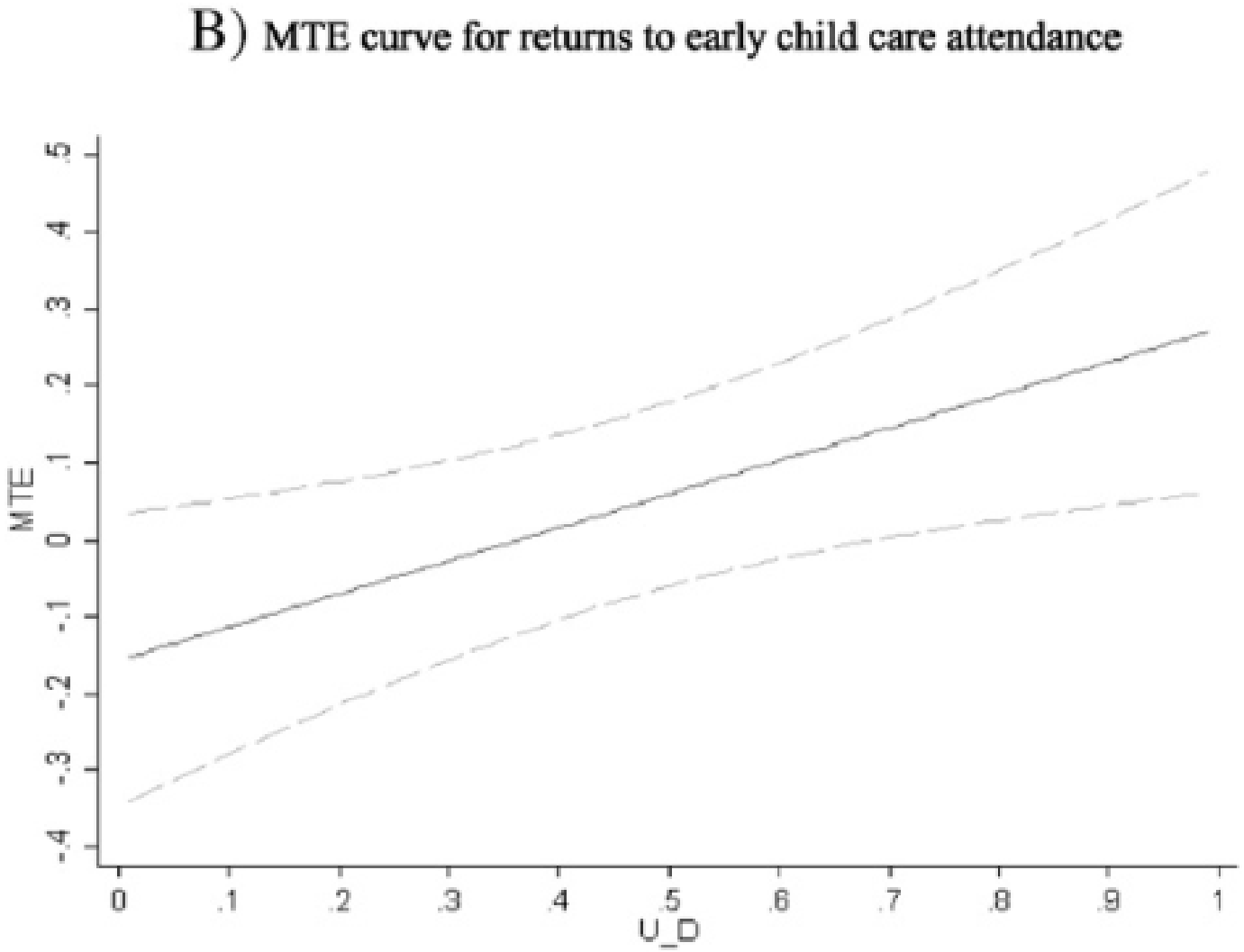
Note: The scale of the *y*-axis is the scale of the MTE, not the scale of the weights, which are scaled to fit the picture.

What does this tell us?

- Huge difference in returns
- Negative selection: people with lowest resistance have returns of 40 percent. For those with highest restistance its a 20 percent loss.
- Suggests people know something we don't when they opt out of college.
- Obviously ATE would be very misleading here.

Selection doesn't have to be positive

Cornellissen 2016: universal pre-K program in Germany.



Summary: Margins Matter

Table 1
Treatment effects parameters

	(1)	(2)
	Returns to college	Returns to early child care attendance
ATE	0.067* (0.038)	0.059 (0.072)
TT	0.143*** (0.035)	−0.051 (0.080)
TUT	−0.007 (0.071)	0.173** (0.085)
IV	0.095** (0.039)	0.065 (0.133)

Notes: The table reports the average treatment effect (ATE), the treatment effect on the treated (TT), treatment effect on the untreated (TUT), and the IV estimate from a linear IV specification for the papers presented in Sections 5.1 and 5.2. Column (1) refers to the results reported in Table 5 in Carneiro et al. (2011). Column (2) refers to the results shown in Table 5, column (1) in Cornelissen et al. (2016). Bootstrapped standard errors are reported in parentheses.

Athey, Susan, Guido W Imbens, Stefan Wager et al. 2016. "Efficient inference of average treatment effects in high dimensions via approximate residual balancing." Tech. rep.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica* 80 (6):2369–2429.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. 2015. "Program evaluation with high-dimensional data." Tech. rep., cemmap working paper, Centre for Microdata Methods and Practice.

Carneiro, Pedro, James J Heckman, and Edward J Vytlacil. 2011. "Estimating marginal returns to education." *American Economic Review* 101 (6):2754–81.

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2016. "From LATE to MTE: Alternative methods for the evaluation of policy interventions." *Labour Economics* 41:47–60.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review* 108 (2):201–40. URL http://www.aeaweb.org/articles?id=10.1257/aer.20161503.

Heckman, James J and Edward Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation 1." *Econometrica* 73 (3):669–738.

Imbens, Guido W. 2015. "Matching methods in practice: Three examples." *Journal of Human Resources* 50 (2):373–419.

King, Gary and Richard Nielsen. Forthcoming. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* .

Lee, David S and Thomas Lemieux. 2010. "Regression discontinuity designs in economics." *Journal of economic literature* 48 (2):281–355.

McCrary, Justin. 2008. "Manipulation of the running variable in the regression discontinuity design: A density test." *Journal of econometrics* 142 (2):698–714.