Treatment Effects

Richard L. Sweeney

based on slides by Chris Conlon

Empirical Methods
Spring 2019

Overview

This lecture draw heavily upon

- 2012 AEA continuing education lectures by Imbens and Woodldridge (full materials available here.)
- Abadie and Cattaneo (2018)

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 (heterogeneity).
    - We don't actually observe any one person's treatment effect ! (Missing Data problem)
    - Individual treatment effects $\tau_i = Y_{1i} - Y_{0i}$ are never observed (FPOCI)
- We need strong assumptions in order to recover $f(\beta_i)$ from data.

More Difficulties

What is hard here?

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

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

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

Structural vs. Reduced Form

- Usually we are interested in one or two parameters of the distribution of β_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. This leads to these as being described as reduced form or quasi-experimental.
- The structural approach attempts to recover the entire joint $f(\beta_i, u_i)$ 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)

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

Confounding

- Consider the *association*

$$\tau = E\left[Y|W=1\right] - E\left[Y|W=0\right]$$

- Then $\tau = \tau_{ATE} + b_{ATE}$
- Where b is the *bias*

$$\begin{aligned} b_{ATE} = &(E\left[Y_1|W=1\right] - E\left[Y_1|W=0\right])Pr(W=0) \\ &+ (E\left[Y_0|W=1\right] - E\left[Y_0|W=1\right])Pr(W=1) \end{aligned}$$

- So the bias disapears only if the potential outcomes are independent of treatment assignment.
- This is called **unconfoundedness**.

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

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

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 I found that matching was helpful in my own work with Julie Mortimer:

- We ran a randomized experiment where we removed Snickers bars from around 60 vending machines in office buildings in downtown Chicago.
- We have 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
6000
4000
2000
0
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 reasearcher 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

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

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.

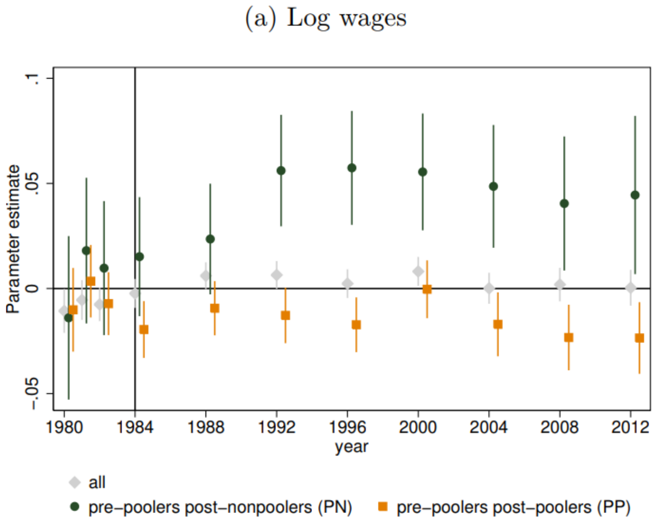
What about matching on the (estimated) propensity score?

- 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

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:

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

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

- Weighted average for people with large π_{1i}.
- Late is treatment effect for those whose probability of treatment is most 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

First stage: Judges Matter

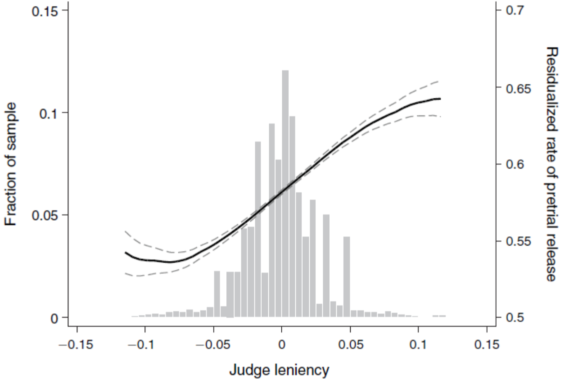


FIGURE 1. DISTRIBUTION OF JUDGE LENIENCY MEASURE AND FIRST STAGE

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:

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

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

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

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)

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

$$= \text{cov}(Z,X)\beta \qquad \text{by (i)}$$

so

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

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:

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

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

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

k > 1: Two stage least squares (TSLS)

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

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

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

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 satistfy 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 instutition
- 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 intepretation.

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

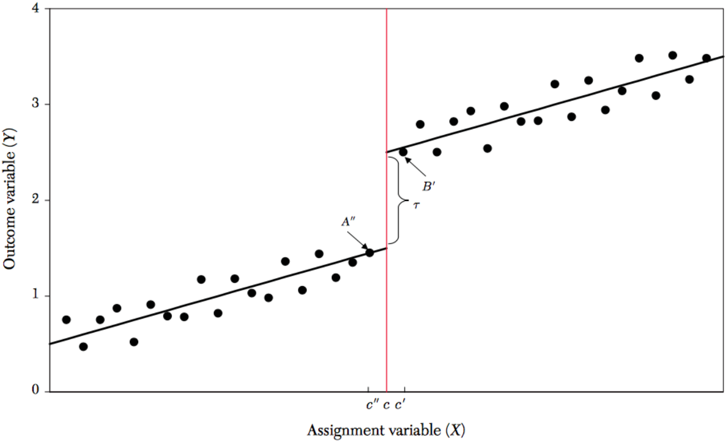
Regression Discontinuity Design

- 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 the JEL Paper by 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

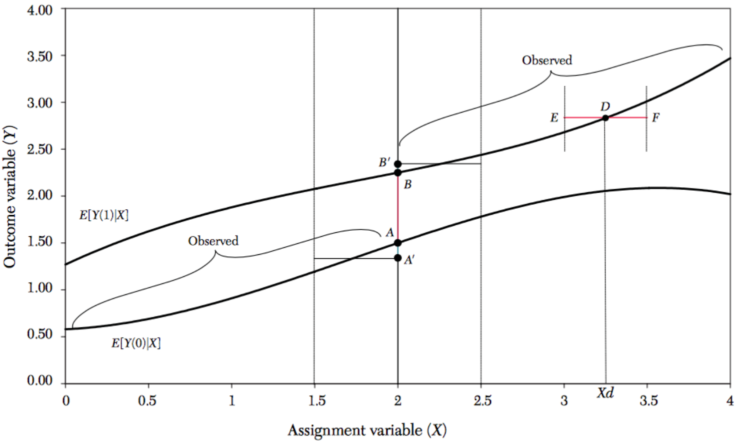
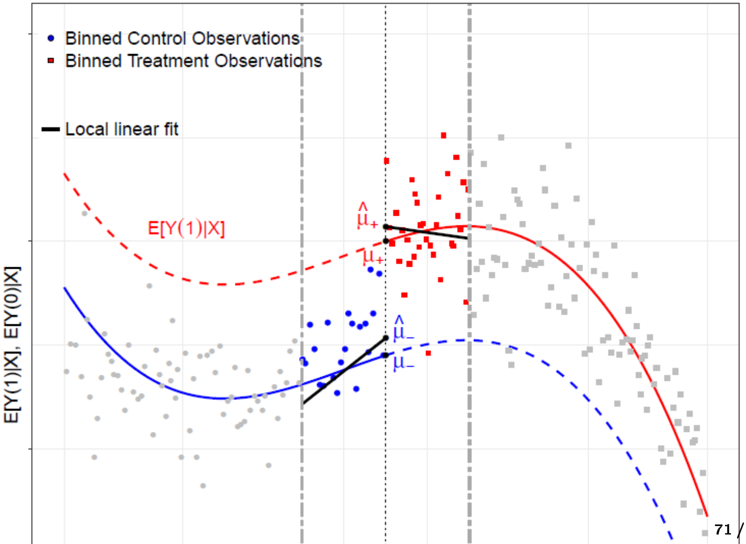


Figure 2. Nonlinear RD

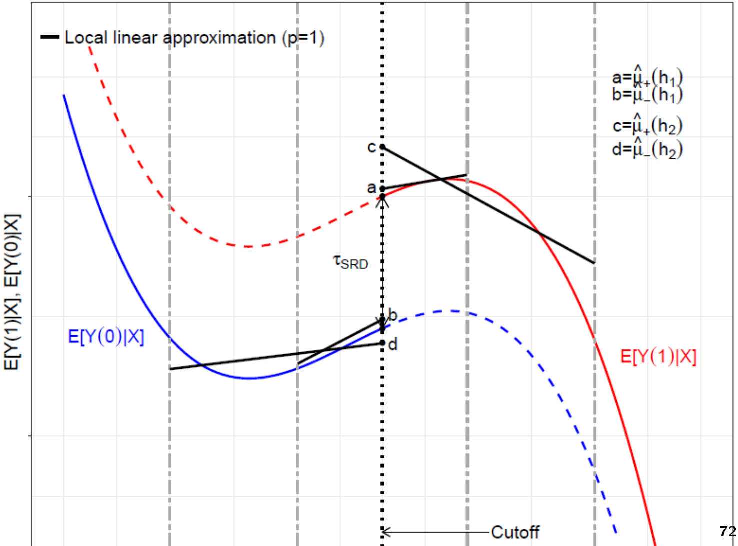
Important reminder: LOCAL effect

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.

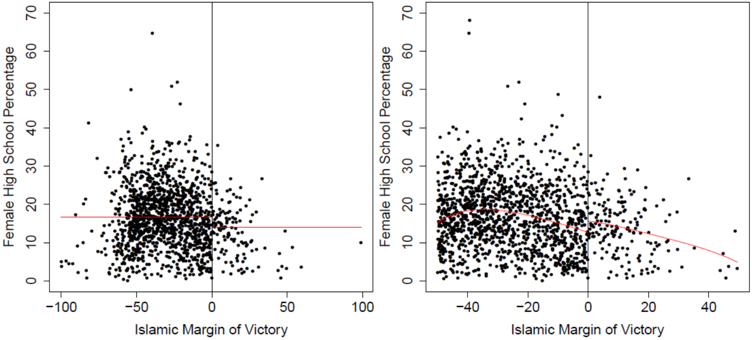
Read Lee and Lemieux (2010) before you get started.

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)

Raw vs Local Comparisons

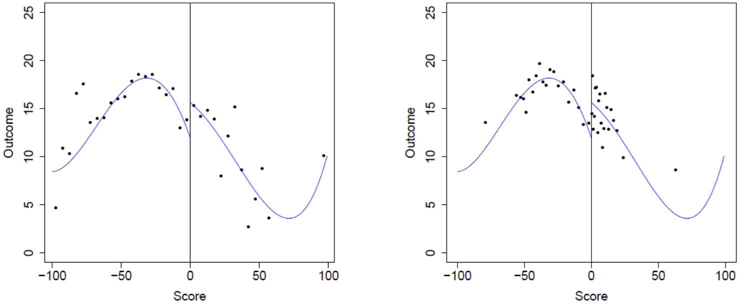
Figure 2.3: Municipalities with Islamic Mayor vs. Municipalities with Secular Mayor—Meyersson data



(a) Raw Comparison of Means

(b) Local Comparison of Means

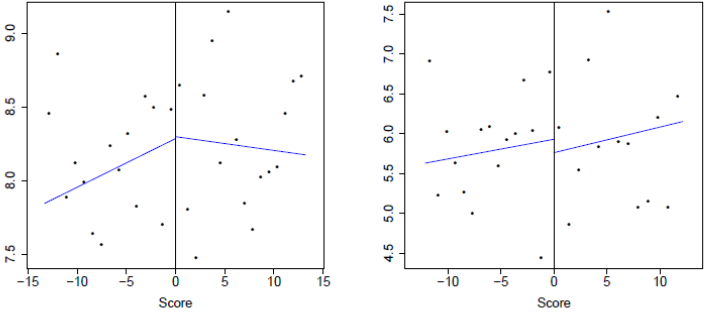
Typically present bincatter



(a) 40 Evenly-Spaced Bins (b) 40 Quantile-Spaced Bins

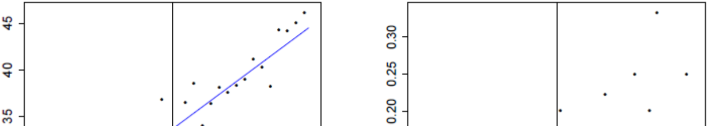
Show other covariates smooth at cutoff

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



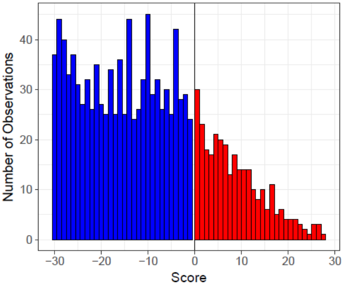
(a) Log Population in 1994

(b) Number of Parties Receiving Votes in 1994

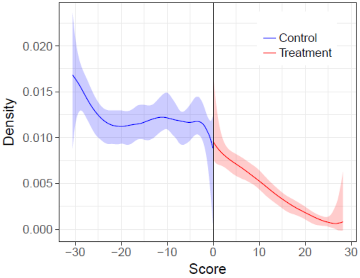


Look for bunching

Figure 5.4: Histogram and Estimated Density of the Score



(a) Histogram

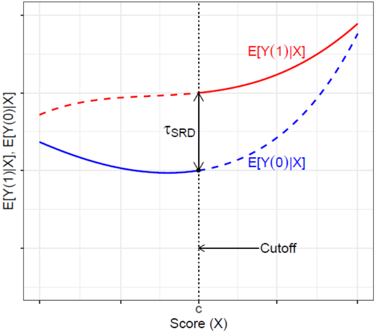


(b) Estimated Density

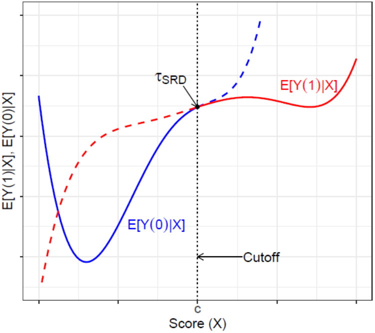
How useful is this LATE?

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.

Difference in Differences

- Sometimes we may feel we can impose more structure on the problem.
- Suppose in particular that we can write the outcome equation as

$$Y_{it} = \alpha_i + d_t + \beta_i T_{it} + u_{it}$$

- In the above we have now introduced a time dimension $t = \{1, 2\}$.
- Now suppose that $T_{i1} = 0$ for all i and $T_{i2} = 1$ for a well defined group of individuals in our population.
- This framework allows us to identify the ATT effect under the assumption that the growth of the outcome in the non-treatment state is independent of treatment allocation:

$$E[Y_{i2}^0 - Y_{i1}^0 | T] = E[Y_{i2}^0 - Y_{i1}^0]$$

- This is known as parallel trends.

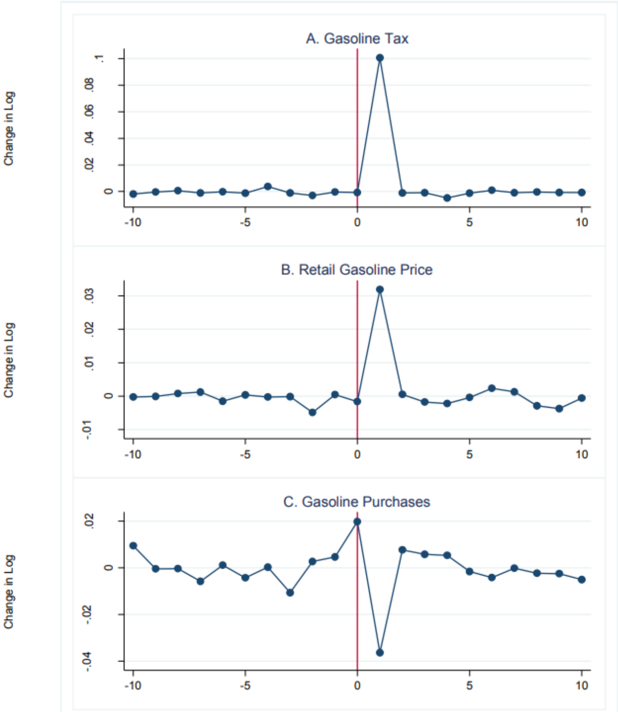
Before and After

An even simpler estimator is the event study.

- We look an outcome before or after an event
  - A news event: the announcement of a merger or stock split.
  - A tax change, a new law, etc.

$$\begin{aligned} E[Y_{i2} - Y_{i1} | T_{i2} = 1] &= E[Y_{i2}^1 - Y_{i1}^1 | T_{i2} = 1] \\ &= d_2 - d_1 + E[\beta_i | T_{i2} = 1] \end{aligned}$$

- Except under strong conditions $d_2 = d_1$ we shouldn't believe the results of the before and after estimator.
- Main Problem: we attribute changes to treatment that might have happened anyway trend.
- e.g: Cigarette consumption drops 4% after a tax hike. (But it dropped 3% the previous four years).
- Also worry about: anticipation, gradual rollout, etc.


A. Gasoline Tax
Change in Log
0
.02
.04
.06
.08
.1
-10
-5
0
5
10
B. Retail Gasoline Price
Change in Log
-.01
0
.01
.02
.03
-10
-5
0
5
10
C. Gasoline Purchases
Change in Log
-.04
-.02
0
.02
-10
-5
0
5
10

Difference in Differences

Let's try and estimate $d_2 - d_1$ directly and then difference it out. Here we use parallel trends:

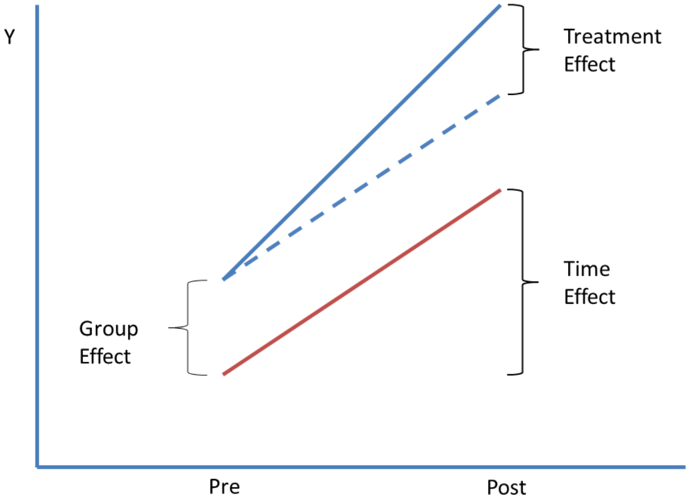
$$E[Y_{i2}^0 - Y_{i1}^0 | T_{i2} = 1] = E[Y_{i2}^0 - Y_{i1}^0 | T_{i2} = 0]$$
$$E[Y_{i2} - Y_{i1} | T_{i2} = 0] = d_2 - d_1$$

We now obtain an estimator for ATT:

$$E[\beta_i | T_{i2} = 1] = E[Y_{i2} - Y_{i1} | T_{i2} = 1] - E[Y_{i2} - Y_{i1} | T_{i2} = 0]$$

which can be estimated by the difference in the growth between the treatment and the control group.

Parallel trends solves a "missing data" problem



Example: Minimum Wage

Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania

By David Card and Alan B. Krueger*

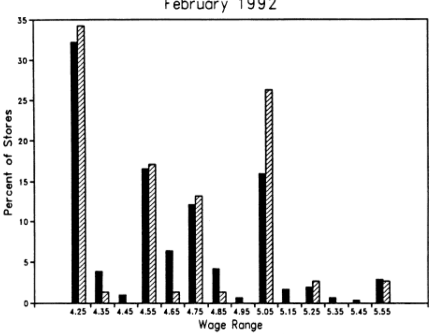
On April 1, 1992, New Jersey's minimum wage rose from $4.25 to $5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above $5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment. (*JEL* J30, J23)

Example: Minimum Wage

TABLE 1—SAMPLE DESIGN AND RESPONSE RATES

	All	Stores in: NJ	Stores in: PA
Wave 1, February 15 – March 4, 1992:			
Number of stores in sample frame:[a]	473	364	109
Number of refusals:	63	33	30
Number interviewed:	410	331	79
Response rate (percentage):	86.7	90.9	72.5
Wave 2, November 5 – December 31, 1992:			
Number of stores in sample frame:	410	331	79
Number closed:	6	5	1
Number under rennovation:	2	2	0
Number temporarily closed:[b]	2	2	0
Number of refusals:	1	1	0
Number interviewed:[c]	399	321	78

[a]Stores with working phone numbers only; 29 stores in original sample frame had disconnected phone numbers.
[b]Includes one store closed because of highway construction and one store closed because of a fire.
[c]Includes 371 phone interviews and 28 personal interviews of stores that refused an initial request for a phone interview.


February 1992
Percent of Stores
Wage Range
4.25 4.35 4.45 4.55 4.65 4.75 4.85 4.95 5.05 5.15 5.25 5.35 5.45 5.55

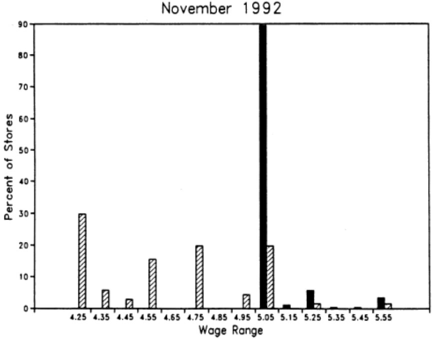

November 1992
Percent of Stores
Wage Range
4.25 4.35 4.45 4.55 4.65 4.75 4.85 4.95 5.05 5.15 5.25 5.35 5.45 5.55

TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE IN NEW JERSEY MINIMUM WAGE

Variable	Stores by state			Stores in New Jersey[a]			Differences within NJ[b]	
	PA (i)	NJ (ii)	Difference, NJ − PA (iii)	Wage = \$4.25 (iv)	Wage = \$4.26–\$4.99 (v)	Wage ≥ \$5.00 (vi)	Low–high (vii)	Midrange–high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	−2.69 (1.37)	−2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	−2.04 (1.14)	3.36 (1.48)	2.91 (1.41)
4. Change in mean FTE employment, balanced sample of stores[c]	−2.28 (1.25)	0.47 (0.48)	2.75 (1.34)	1.21 (0.82)	0.71 (0.69)	−2.16 (1.01)	3.36 (1.30)	2.87 (1.22)
5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0[d]	−2.28 (1.25)	0.23 (0.49)	2.51 (1.35)	0.90 (0.87)	0.49 (0.69)	−2.39 (1.02)	3.29 (1.34)	2.88 (1.23)

Notes: Standard errors are shown in parentheses. The sample consists of all stores with available data on employment. FTE (full-time-equivalent) employment counts each part-time worker as half a full-time worker. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing.

[a]Stores in New Jersey were classified by whether starting wage in wave 1 equals \$4.25 per hour ($N = 101$), is between \$4.26 and \$4.99 per hour ($N = 140$), or is \$5.00 per hour or higher ($N = 73$).

[b]Difference in employment between low-wage (\$4.25 per hour) and high-wage (≥ \$5.00 per hour) stores; and difference in employment between midrange (\$4.26–\$4.99 per hour) and high-wage stores.

[c]Subset of stores with available employment data in wave 1 and wave 2.

[d]In this row only, wave-2 employment at four temporarily closed stores is set to 0. Employment changes are based on the subset of stores with available employment data in wave 1 and wave 2.

Difference in Differences

Now consider the following problem:

- Suppose we wish to evaluate a training program for those with low earnings. Let the threshold for eligibility be B.
- We have a panel of individuals and those with low earnings qualify for training, forming the treatment group.
- Those with higher earnings form the control group.
- Now the low earning group is low for two reasons
  1. They have low permanent earnings (α_i is low) - this is accounted for by diff in diffs.
  2. They have a negative transitory shock (u_{i1} is low) - this is not accounted for by diff in diffs.
- #2 above violates the assumption $E[Y_{i2}^0 - Y_{i1}^0|T] = E[Y_{i2}^0 - Y_{i1}^0]$.
- This is effectively regression to the mean: those unlucky enough to have a bad shock recover and show greater growth relative to those with a good shock. The nature of the bias depends on the stochastic properties of the shocks and how individuals select into training.

Who get's treated?

- The assumption on growth of the non-treatment outcome being independent of assignment to treatment may be violated, but it may still be true conditional on X.
- Consider the assumption

$$E[Y_{i2}^0 - Y_{i1}^0 | X, T] = E[Y_{i2}^0 - Y_{i1}^0 | X]$$

- This is just matching assumption on a redefined variable, namely the growth in the outcomes. In its simplest form the approach is implemented by running the regression

$$Y_{it} = \alpha_i + d_t + \beta_i T_{it} + \gamma_t' X_i + u_{it}$$

which allows for differential trends in the non-treatment growth depending on X_i. More generally one can implement propensity score matching on the growth of outcome variable when panel data is available.

DiD with Repeated Cross Sections

- Suppose we do not have available panel data but just a random sample from the relevant population in a pre-treatment and a post-treatment period.
- First consider a simple case where $E[Y^0_{i2} - Y^0_{i1}|T] = E[Y^0_{i2} - Y^0_{i1}]$.
- We need to modify slightly the assumption to

$$E[Y^0_{i2}|_{\text{Group receiving training}}] - E[Y^0_{i1}|_{\text{Group receiving training in the next period}}] = E[Y^0_{i2} - Y^0_{i1}]$$

  which requires additional assumption that the population we will be sampling from does not change composition.
- We can then obtain immediately an estimator for ATT as

$$E[\beta_i|T_{i2} = 1]$$
$$= E[Y_{i2}|_{\text{Group receiving training}}] - E[Y_{i1}|_{\text{Group receiving training next period}}]$$
$$- \{E[Y_{i2}|_{\text{Non-trainees}}] - E[Y_{i1}|_{\text{Group not receiving training next period}}]\}$$

Difference in Differences with Repeated Cross Sections

- More generally we need an assumption of conditional independence of the form

$$E[Y_{i2}^0|X, \text{Group receiving training}] - E[Y_{i1}^0|X, \text{Group receiving training next period}]$$
$$= E[Y_{i2}^0|X] - E[Y_{i1}^0|X]$$

- Under this assumption (and some auxiliary parametric assumptions) we can obtain an estimate of the effect of treatment on the treated by the regression

$$Y_{it} = \alpha_g + d_t + \beta T_{it} + \gamma' X_{it} + u_{it}$$

Difference in Differences with Repeated Cross Sections

- More generally we can first run the regression

$$Y_{it} = \alpha_g + d_t + \beta(X_{it})T_{it} + \gamma' X_{it} + u_{it}$$

where α_g is a dummy for the treatment of comparison group, and $\beta(X_{it})$ can be parameterized as $\beta(X_{it}) = \beta' X_{it}$. The ATT can then be estimated as the average of $\beta' X_{it}$ over the (empirical) distribution of X.

- A non parametric alternative is offered by Blundell, Dias, Meghir and van Reenen (2004).

DiD vs Fixed Effects

- What if we have a long panel with many similar changes?
  - Greenstone (2002): Counties move in and out of Clean Air Act
  - Evans, Ringel, and Stech (1999): Since 1975, more than 200 state cigarette tax changes
- Fixed effects generalize DD with $T > 2$ periods and $J > 2$ groups
- Advantage relative to DD: more precise estimates by pooling several changes
- Disadvantage: fixed effects is a black-box regression, more difficult to check trends non-parametrically as with a single change

The best DiD's can be seen graphically

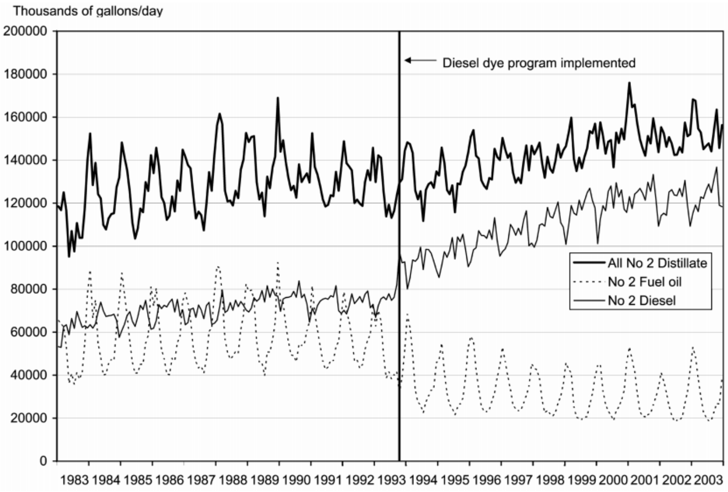


Fig. 2.—U.S. sales of No. 2 distillate

What about triple differencing?

- Sometime we might use a "placebo" DD to make parallel trends more convincing
- Example: Imagine a policy which offered STEM outreach to high school girls in Massachusetts
  - Natural DiD control group: boys in MA
  - However over time there could be general shifts in the relative outcomes of boys and girls everywhere
  - Suggest looking at how the difference between boys and girls in MA changed relative to the changes in other states (say RI)
- Logically sound, but much harder to see/ validate visually

Difference in Differences and Selection on Unobservables

- Suppose we relax the assumption of *no selection* on unobservables.
- Instead we can start by assuming that

$$E[Y_{i2}^0|X,Z] - E[Y_{i1}^0|X,Z] = E[Y_{i2}^0|X] - E[Y_{i1}^0|X]$$

  where Z is an instrument which determines training eligibility say but does not determine outcomes in the non-training state. Take Z as binary (1,0).
- Non-Compliance: not all members of the eligible group $(Z = 1)$ will take up training and some of those ineligible $(Z = 0)$ may obtain training by other means.
- A difference in differences approach based on grouping by Z will estimate the impact of being allocated to the eligible group, but not the impact of training itself.

Difference in Differences and Selection on Unobservables

- Now suppose we still wish to estimate the impact of training on those being trained (rather than just the effect of being eligible)
- This becomes an IV problem and following up from the discussion of LATE we need stronger assumptions
    - Independence: for $Z = a, \{Y_{i2}^0 - Y_{i1}^0, Y_{i2}^1 - Y_{i1}^1, T(Z = a)\}$ is independent of Z.
    - Monotonicity $T_i(1) \geq T_i(0) \, \forall \, i$
- In this case LATE is defined by

$$[E(\Delta Y|Z = 1) - E(\Delta Y|Z = 0)]/[Pr(T(1) = 1) - Pr(T(0) =$$

assuming that the probability of training in the first period is zero.

Synthetic Controls

- DiD methods compare two groups before and after some change.
- Challenge: What's a good comparison group? Even if you pick the best available option, might not track eachother that closely even in the pre-period.
- Moreover, if we don't have another untreated group that is well balanced against the treatment group, are we stuck?
- Synthetic control methods pick weighted averages from control population to construct better comparisons (Abadie and Gardeazabal, 2003; Abadie, Diamond, and Hainmueller, 2010)
- Athey and Imbens (2017) call this "arguably the most important innovation in the policy evaluation literature in the past 15 years".

Initial motivation: Case studies

- Often we're interested in the aggregate effects of large, singular policies.
  - What was the impact of MassHealth?
  - Fukushima
  - Terrorism
  - German Re-unification
- What would a rigorous "case study" of these look like?

ADH (JASA 2010)

- Consider a panel with $J+1$ units observed for $t = 1, 2, ..., T$ periods.
- Unit 1 exposed to treatment in period T_0 (continues to T)
- Synthetic control estimator is

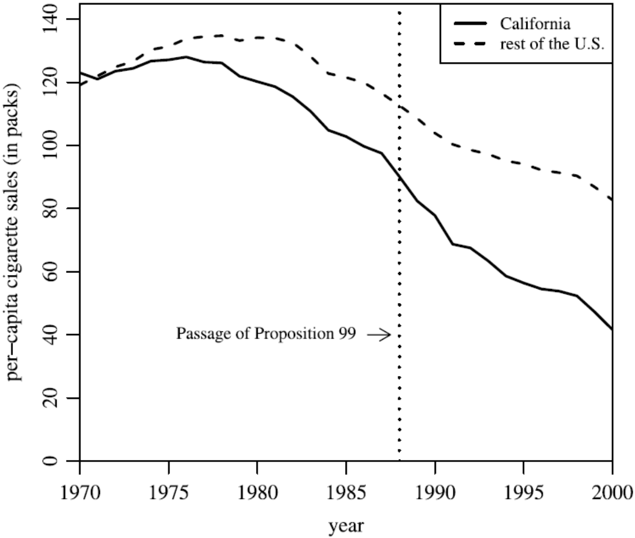
$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

  where w is a collection of weights.
- In Abadie, Diamond, and Hainmueller (2010) the (non-negative) weights are chosen to minimize the distance between some chosen vector of preintervention characteristics (and sum to one).
- Subsequent literature has relaxed these.

ADH Example: CA Prop 99

- Anti cigarette law in CA in 1988
    - increased state excise tax by 25 cents per pack
    - earmarked the tax revenues to health and anti-smoking education budgets
    - funded anti-smoking media campaigns
    - spurred local clean indoor-air ordinances throughout the state
- What was the net effect on sales?

Sales were trending down everywhere



What does synthetic CA look like?

Table 2. State weights in the synthetic California

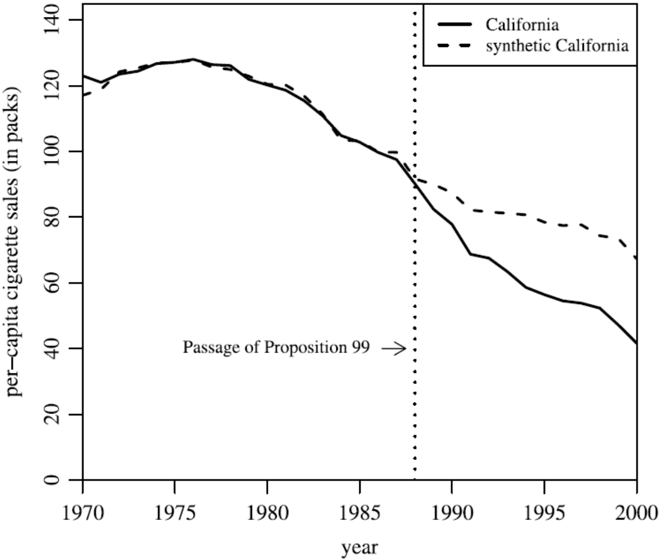
State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

Balance Acheived

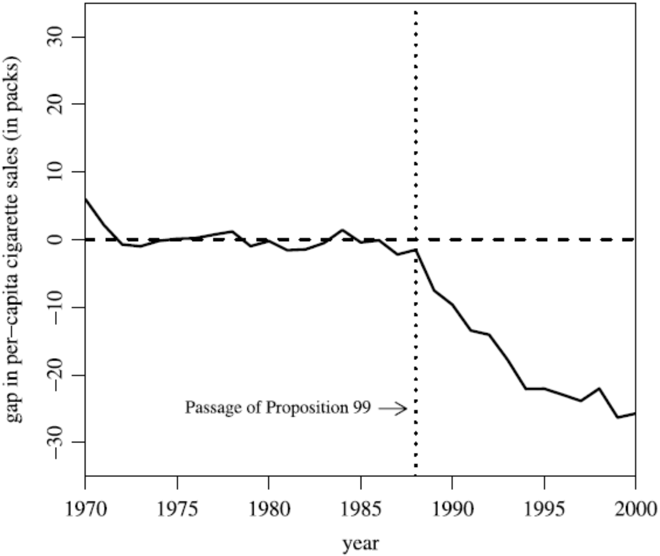
Table 1. Cigarette sales predictor means

Variables	California Real	California Synthetic	Average of 38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

NOTE: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988). GDP per capita is measured in 1997 dollars, retail prices are measured in cents, beer consumption is measured in gallons, and cigarette sales are measured in packs.


California
synthetic California
Passage of Proposition 99
per–capita cigarette sales (in packs)
year
0
20
40
60
80
100
120
140
1970
1975
1980
1985
1990
1995
2000

Parallel trends acheived by contstruction



What about inference

- SE's typically reported reflect uncertainty in sample relative to aggregate population.
- ADH propose using a placebo test to assess null of no change in CA.
- Steps:
  1. Randomly select one of the other J control units / time cutoffs and declare it treated.
  2. Construct synthetic controls and estimate ATT.
  3. Repeat many times
- Since none of these units are actually treated, this test distribution simulates distribution of the differences relative to the synthetic control under the true null of no effect.

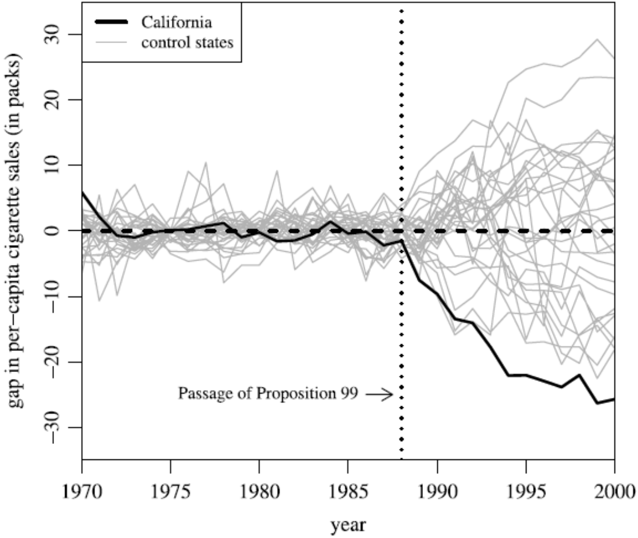


Figure 6. Per-capita cigarette sales gaps in California and placebo gaps in 29 control states (discards states with pre-Proposition 99 MSPE five times higher than California's).

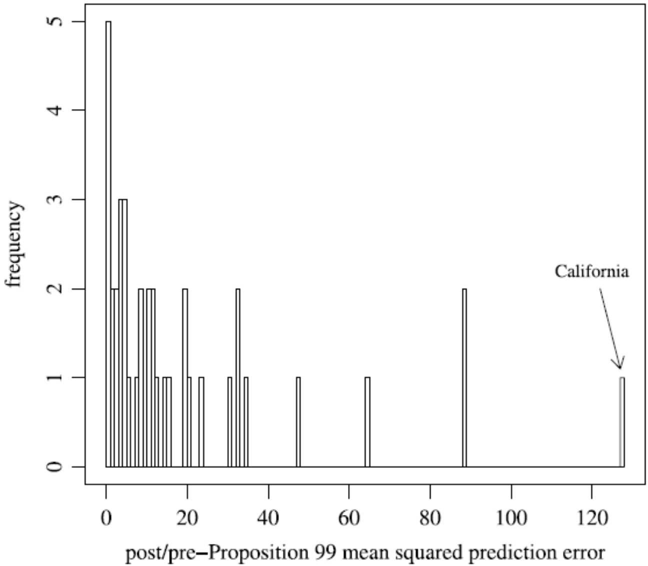


Figure 8. Ratio of post-Proposition 99 MSPE and pre-Proposition 99 MSPE: California and 38 control states.

A synthesis of approaches

- Doudchenko and Imbens (2016) attempt to synthesize several of the approaches discussed so far balancing (matching), regression, DiD, synthetic controls.
- These methods can be recast as trying to impute the untreated outcome for treated unit 0 to estimate $\hat{\tau}_{0,T} = Y_{0,T}(1) - \hat{Y}_{0,T}(0)$
- many impose the linear structure

$$\hat{Y}_{0,T}(0) = \mu + \sum_{i}^{N} w_i \dot{Y}_{i,T}^{obs}$$

DI consider a series of contstraints

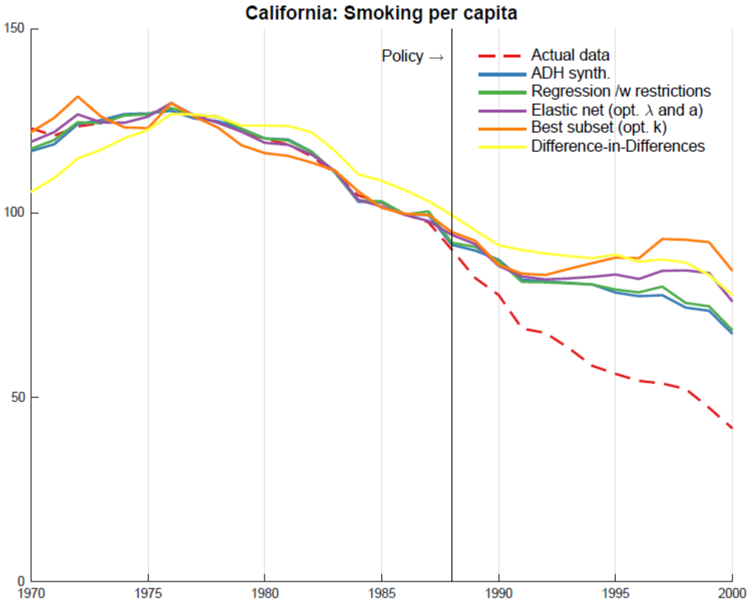
1. no intercept: $\mu = 0$
2. adding up: $\sum_i^N w_i = 1$
3. non-negativity: $w_i > 0$
4. exact balancing: $Y_{t,pre}^{obs} = \mu + w^T \mathbf{Y_{c,pre}^{obs}}$
5. constant weights: $w_i = \bar{w}$

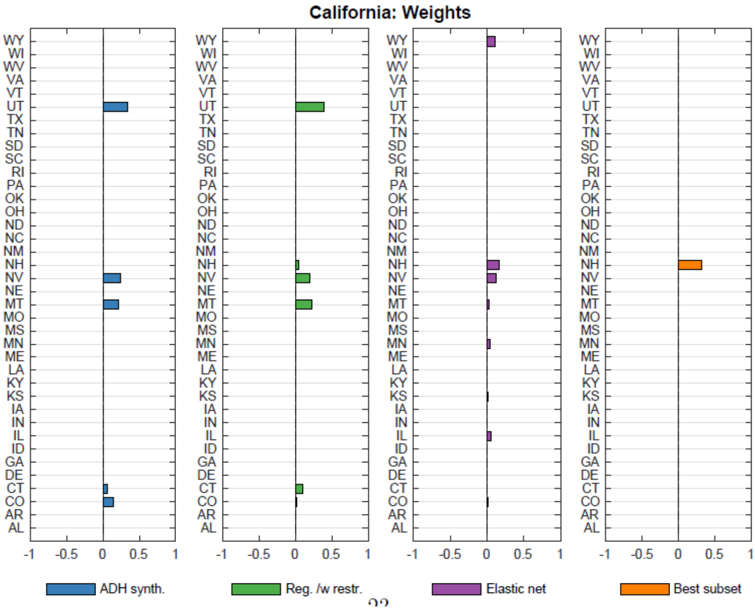
- Assumptions 1-3 are imposed by ADH
- No intercept actually precludes defining feature of DiD
- Contstant weights implicit when number of control units is large
- Other useful features of an objective function:
  - match pre period *outcomes* well
  - small number of features
  - non-disperse set of parameters
- All of these generally suggest regularization.

this paper. These five estimators include the original ADH estimator, the constrained estimator with the same restrictions, $\mu = 0$, $\sum_{i=1}^{N} \omega_i = 1$ and $\omega_i \geq 0$, the best subset estimator, and DID estimator, and the elastic net estimator. For the best subset estimator the optimal number of controls, based on cross-validation, is 1. For the elastic net estimator the tuning parameters, choosen by cross-validation, are $\alpha = 0.1$ and $\lambda = 45.5$, leading to 8 states with non-zero weights, all of them positive.

Table 1: California: Parameters

Model	$\sum_i w_i$	α
Original synth.	1	0
Constrained reg.	1	0
Elastic net	0.55	18.5
Best subset	0.32	37.6
Diff-in-diff	1	−14.4


California: Smoking per capita
Policy →
Actual data
ADH synth.
Regression /w restrictions
Elastic net (opt. λ and a)
Best subset (opt. k)
Difference-in-Differences
150
100
50
0
1970
1975
1980
1985
1990
1995
2000


California: Weights
WY
WI
WV
VA
VT
UT
TX
TN
SD
SC
RI
PA
OK
OH
ND
NC
NM
NH
NV
NE
MT
MO
MS
MN
ME
LA
KY
KS
IA
IN
IL
ID
GA
DE
CT
CO
AR
AL
-1
-0.5
0
0.5
1
ADH synth.
Reg. /w restr.
Elastic net
Best subset

One quantity to rule them all: MTE

Heckman and Vytlacil provide a unifying non-parametric framework to categorize treatment effects. Their approach is known as the marginal treatment effect or MTE

- The MTE isn't a number it is a function.
- 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 $f(\beta_i)$.

One quantity to rule them all: MTE

- Consider a treatment effect $\beta_i = Y_i(1) - Y_i(0)$.
- Think about a single-index such that $T_i = 1(v_i \leq Z_i'\gamma)$.
- Think about the person for whom $v_i = Z_i'\gamma$ (just barely untreated).

$$\Delta^{MTE}(X_i, v_i) = E[\beta_i | X_i, v_i = Z_i'\gamma]$$

- MTE is average impact of receiving a treatment for everyone with the same $Z'\gamma$.
- For any single index model we can rewrite

$$T_i = 1(v_i \leq Z_i'\gamma) = 1(u_{is} \leq F(Z_i'\gamma)) \text{ for } u_s \in [0,1]$$

- F is just the cdf of v_i
- Now we can write $P(Z) = Pr(T = 1|Z) = F(Z'\gamma)$.

MTE: Derivation

Now we can write,

$$\begin{aligned} Y_0 &= \gamma_0' X + U_0 \\ Y_1 &= \gamma_1' X + U_1 \end{aligned}$$

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

1. $(U_0, U_1, u_s) \perp P(Z)|X$. (Exogeneity)
2. 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. Much more like the special regressor method we will discus next time.

MTE: Derivation

Now we can write,

$$
\begin{aligned}
Y &= \gamma_0' X + T(\gamma_1 - \gamma_0)' X + U_0 + T(U_1 - U_0) \\
E[Y|X, P(Z) = p] &= \gamma_0' X + p(\gamma_1 - \gamma_0)' X + E[T(U_1 - U_0)|X, P(Z) = p]
\end{aligned}
$$

Observe $T = 1$ over the interval $u_s = [0, p]$ and zero for higher values of u_s. Let $U_1 - U_0 \equiv \eta$.

$$
\begin{aligned}
E[T(U_1 - U_0)|P(Z) = p, X] &= \int_{-\infty}^{\infty} \int_0^p (U_1 - U_0) f((U_1 - U_0)|U_s = u_s) d \\
E[T(\eta)|P(Z) = p, X] &= \int_{-\infty}^{\infty} \int_0^p \eta f(\eta|U_s = u_s) d\eta d u_s
\end{aligned}
$$

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

What is $E[\eta|u_s = p]$? The expected unobserved gain from treatment of those people who are on the treatment/no-treatment margin $P(Z) = p$.

How to Estimate an MTE

Easy

1. Estimate $P(Z) = Pr(T = 1|Z)$ nonparametrically (include exogenous part of X in Z).
2. Nonparametric regression of Y on X and $P(Z)$ (polynomials?)
3. Differentiate w.r.t. $P(Z)$
4. 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 $\Delta^{MTE}(p)$.

Everything is an MTE

Calculate the outcome given (X, Z) (actually X and $P(Z) = p$).

- ATE : This one is obvious. We treat everyone!

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

- LATE: Fix an X and $P(Z)$ varies from $b(X)$ to $a(X)$ and we integrated over the area between (compliers).

$$LATE(X) = \int_{-\infty}^{\infty} \Delta^{MTE}(p) = (\gamma_1 - \gamma_0)'X + \frac{1}{a(X) - b(X)} \int_{b(X)}^{a(X)} E(\eta$$

- ATT

$$TT(X) = \int_{-\infty}^{\infty} \Delta^{MTE}(p) \frac{Pr(P(Z|X) > p)}{E[P(Z|X)]} d\,p$$

- Weights for IV and OLS are a bit more complicated. See the Heckman and Vytlacil paper(s).

Carneiro, Heckman and Vytlacil (AER 2010)

- Estimate returns to college (including heterogeneity of returns).
- NLSY 1979
- $Y = \log(wage)$
- Covariates X: Experience (years), Ability (AFQT Score), Mother's Education, Cohort Dummies, State Unemployment, MSA level average wage.
- Instruments Z: College in MSA at age 14, average earnings in MSA at 17 (opportunity cost), avg unemployment rate in state.

Carneiro, Heckman and Vytlacil

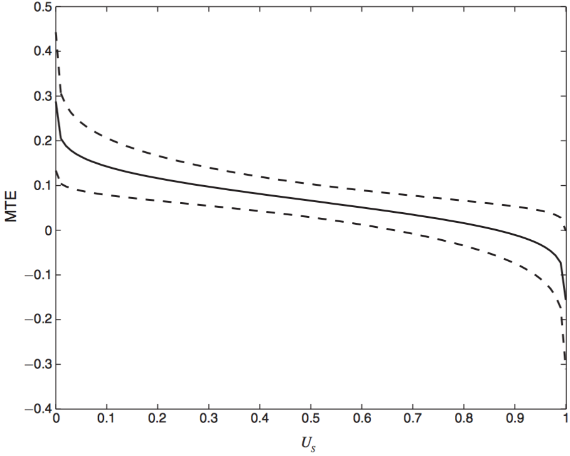


FIGURE 1. MTE ESTIMATED FROM A NORMAL SELECTION MODEL

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:

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

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			

Carneiro, Heckman and Vytlacil

TABLE 5—RETURNS TO A YEAR OF COLLEGE

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

Carneiro, Heckman and Vytlacil

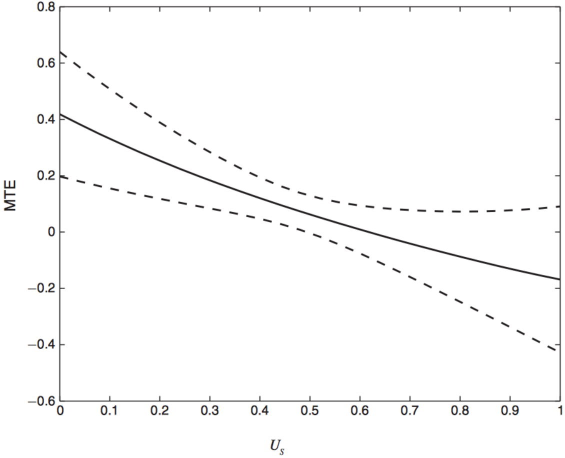


FIGURE 4. $E(Y_1 - Y_0 | \mathbf{X}, U_S)$ WITH 90 PERCENT CONFIDENCE INTERVAL—LOCALLY QUADRATIC REGRESSION ESTIMATES

Notes: To estimate the function plotted here, we first use a partially linear regression of log wages on polynomials in $\mathbf{X}$, interactions of polynomials in $\mathbf{X}$ 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; $\mathbf{X}$ includes experience, current average earnings in the

Carneiro, Heckman and Vytlacil

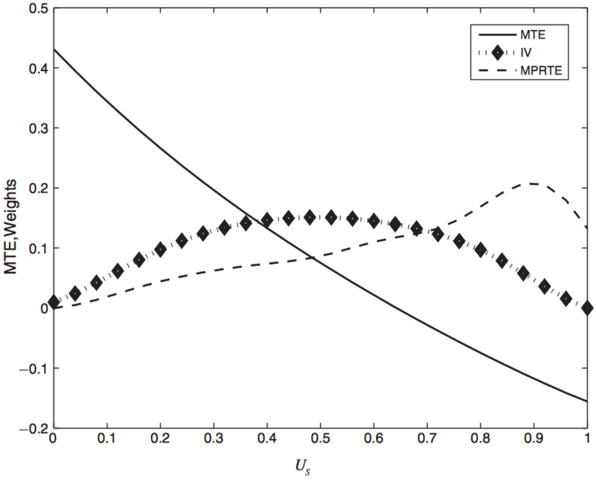


Figure 6. Weights for IV and MPRTE

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.

Diversion Example

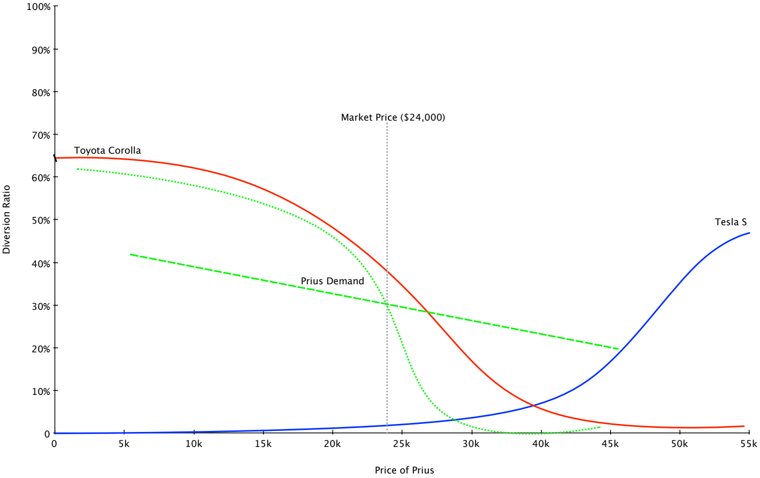
I have done some work trying to bring these methods into merger analysis.

- Key quantity: Diversion Ratio as I raise my price, how much do people switch to a particular competitor's product

$$D_{jk}(p_j, p_{-j}) = \left| \frac{\partial q_k}{\partial p_j}(p_j, p_{-j}) / \frac{\partial q_j}{\partial p_j}(p_j, p_{-j}) \right|$$

- We hold p_{-j} fixed and trace out $D_{jk}(p_j)$.
- The treatment is leaving good j.
- The Y_i is increased sales of good k.
- The Z_i is the price of good j.
- The key is that all changes in sales of k come through people leaving good j (no direct effects).

Diversion for Prius (FAKE!)



Diversion Example

$$\widehat{D_{jk}^{LATE}} = \frac{1}{\Delta q_j} \int_{p_j^0}^{p_j^0+\Delta p_j} \underbrace{\frac{\partial q_k(p_j, p_{-j}^0)}{\partial q_j}}_{\equiv D_{jk}(p_j, p_{-j}^0)} \left| \frac{\partial q_j(p_j, p_{-j}^0)}{\partial p_j} \right| dp_j$$

- $D_{jk}(p_j, p_{-j}^0)$ is the MTE.
- Weights $w(p_j) = \frac{1}{\Delta q_j} \frac{\partial q_j(p_j, p_{-j}^0)}{\partial p_j}$ correspond to the lost sales of j at a particular p_j as a fraction of all lost sales.
- When is $LATE \approx ATE$?
  - Demand for Prius is steep: everyone leaves right away
  - $D_{j,k}(p_j)$ is relatively flat.
  - We might want to think about raising the price to choke price (or eliminating the product from the consumers choice set) same as treating everyone!

Abadie, Alberto and Matias D. Cattaneo. 2018. "Econometric Methods for Program Evaluation." *Annual Review of Economics* 10 (1):465–503. URL https://doi.org/10.1146/annurev-economics-080217-053402.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American statistical Association* 105 (490):493–505.

Abadie, Alberto and Javier Gardeazabal. 2003. "The economic costs of conflict: A case study of the Basque Country." *American economic review* 93 (1):113–132.

Athey, Susan and Guido W Imbens. 2017. "The state of applied econometrics: Causality and policy evaluation." *Journal of Economic Perspectives* 31 (2):3–32.

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

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.

Doudchenko, Nikolay and Guido W Imbens. 2016. "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis." Working Paper 22791, National Bureau of Economic Research. URL http://www.nber.org/papers/w22791.

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* .