Part 8: Treatment Effects

Chris Conlon

March 23, 2020

Applied Econometrics

Intro

Overview

This Lecture will cover (roughly) the following papers:

Theory:

- Angrist and Imbens (1994)
- Heckman Vytlacil (2005/2007)
- Abadie and Imbens (2006)

And draw heavily upon notes by

- Guido Imbens
- Richard Blundell and Costas Meghir

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

Example: Borjas (1987)

- Consider two countries $(0/1)$ (source and host).

$$\begin{aligned} \ln w_0 &= \alpha_0 + u_0 \quad \text{with } u_0 \sim N(0, \sigma_0^2) \text{ source country} \\ \ln w_1 &= \alpha_1 + u_1 \quad \text{with } u_1 \sim N(0, \sigma_1^2) \text{ host country} \end{aligned}$$

- Now we allow for migration cost of C which he writes in hours: $\pi = \frac{C}{w_0}$.
- Assume workers know everything; you only see u_0 OR u_1 depending on country.
- Correlation in earnings is $\rho = \frac{\sigma_{01}}{\sigma_0 \sigma_1}$.

Example: Borjas (1987)

- Workers will migrate if:

$$(\alpha_1 - \alpha_0 - \pi) + (u_1 - u_0) > 0$$

- Who migrates? Probability of migration. Define $\nu = u_1 - u_0$.

$$\begin{aligned} P = \Pr\left[\nu > (\alpha_0 - \alpha_1 + \pi)\right] &= \Pr\left[\frac{\nu}{\sigma_\nu} > \frac{(\alpha_0 - \alpha_1 + \pi)}{\sigma_\nu}\right] \\ &= 1 - \Phi\left(\frac{(\alpha_0 - \alpha_1 + \pi)}{\sigma_\nu}\right) \equiv 1 - \Phi(z) \end{aligned}$$

- Higher $z \rightarrow$ less migration.

Example: Borjas (1987): How does selection work?

Construct counterfactual wages for workers in source country for those who immigrate:

- For now ignore mean differences $\alpha_0 = \alpha_1 = \alpha$.

$$\begin{aligned} E\left(w_0 \middle| \text{ Immigrate }\right) &= \alpha + E\left(\varepsilon_0 \middle| \frac{\nu}{\sigma_\nu} > z\right) \\ &= \alpha + \sigma_0 E\left(\frac{\varepsilon_0}{\sigma_0} \middle| \frac{\nu}{\sigma_\nu} > z\right) \end{aligned}$$

- Wages depend on:
  1. Mean earnings in the source country
  2. Both error terms (u_0, u_1) through ν
  3. Implicitly, it also depends on the correlation between the error terms.

Example: Borjas (1987): How does selection work?

- If everything is normal, we just run univariate regression $E\left(u_0|\nu\right) = \frac{\sigma_{0\nu}}{\sigma_\nu^2}\nu$:

$$E\left(\frac{u_0}{\sigma_0}|\frac{\nu}{\sigma_\nu}\right) = \frac{1}{\sigma_0}\cdot\frac{\sigma_{0\nu}}{\sigma_\nu^2}\cdot\frac{\sigma_\nu^2}{\sigma_\nu^2}\cdot\nu = \frac{\sigma_{0\nu}}{\sigma_0\sigma_\nu}\frac{\nu}{\sigma_\nu} = \rho_{0\nu}\frac{\nu}{\sigma_\nu}$$

$$\begin{aligned} E\left(w_0|\text{ Immigrate }\right) &= \alpha_0 + \sigma_0 E\left(\frac{u_0}{\sigma_0}|\frac{\nu}{\sigma_\nu} > z\right) \\ &= \alpha_0 + \rho_{0\nu}\sigma_0 E\left(\frac{\nu}{\sigma_\nu}|\frac{\nu}{\sigma_\nu} > z\right) \\ &= \alpha_0 + \rho_{0\nu}\sigma_0\left(\frac{\phi(z)}{1-\Phi(z)}\right) \end{aligned}$$

- This hazard rate of the standard normal has a special name Inverse Mills Ratio $E[x|x > z]$.

Example: Borjas (1987): How does selection work?

- A similar expression for those who do immigrate:

$$\begin{aligned} E\left(w_1 \middle| \text{ Immigrate } \right) &= \alpha_1 + E\left(u_1 | \frac{\nu}{\sigma_\nu} > z\right) \\ &= \alpha_1 + \rho_{1\nu}\sigma_1 \left(\frac{\phi(z)}{\Phi(-z)}\right) \end{aligned}$$

- We can re-write both expressions in terms of the Inverse Mills Ratio

Inverse Mills Ratio

$$\begin{aligned}
E\left(w_0 \middle| \text{ Immigrate }\right) &= \alpha_0 + \rho_{0\nu}\sigma_0 \left(\frac{\phi(z)}{1-\Phi(z)}\right) \\
&= \alpha_0 + \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\rho - \frac{\sigma_0}{\sigma_1}\right)\left(\frac{\phi(z)}{1-\Phi(z)}\right) \\
E\left(w_1 \middle| \text{ Immigrate }\right) &= \alpha_1 + \rho_{1\nu}\sigma_1 \left(\frac{\phi(z)}{1-\Phi(z)}\right) \\
&= \alpha_1 + \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\frac{\sigma_1}{\sigma_0} - \rho\right)\left(\frac{\phi(z)}{1-\Phi(z)}\right)
\end{aligned}$$

Where $\rho = \sigma_{01}/\sigma_0\sigma_1$.

Positive Hierarchical Sorting

Let $Q_0 = E\left(u_0 | I = 1\right), Q_1 = E\left(u_1 | I = 1\right)$ (expected skill of immigrants).

- Immigrants are positively selected and above average $(Q_0, Q_1) > 0$ and $\frac{\sigma_1}{\sigma_0} > 1$ and $\rho > \frac{\sigma_0}{\sigma_1}$
  - $\frac{\sigma_1}{\sigma_0} > 1$ returns to "skill" are higher in host country.
  - $\rho > \frac{\sigma_0}{\sigma_1}$ correlation between valued skills in both counties is high (similar skills valued in both countries).
- Best and brightest leave because returns to skill are too low in home country.

Negative Hierarchical Sorting

We swap the standard deviations:

- Immigrants are negatively selected and below average $(Q_0, Q_1) < 0$ and $\frac{\sigma_1}{\sigma_0} > 1$ and $\rho > \frac{\sigma_0}{\sigma_1}$
    - $\frac{\sigma_0}{\sigma_1} > 1$ returns to "skill" are lower in host country.
    - $\rho > \frac{\sigma_1}{\sigma_0}$ correlation between valued skills in both counties is high (similar skills valued in both countries).
- Compressed wage structure attracts the low skill types because it provides "insurance" or "subsidizes" low wage workers.

Refugee/Superman Sorting?

- Immigrants are below average at home and above average in host $(Q_0 < 0, Q_1 > 1)$ and $\frac{\sigma_1}{\sigma_0} > 1$:
  - $\rho < \min\left(\frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1}\right)$ being below average in source country makes you above average in host country.
- You are a nerdy intellectual in a country that values physical labor, or are otherwise discriminated against in the labor market.

The missing (fourth) case:

- Mathematically impossible $\rho > \max\left(\frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1}\right)$

The Evaluation Problem

- Define an outcome variable Y_i for each individual
- Two potential outcomes for each person $\{Y_i(1), Y_i(0)\}$ depending on whether they receive treatment or not.
- Call $Y_i(1) - Y_i(0) = \beta_i$ the Treatment effect.
- Two major problems:
  - All individuals have different treatment effects (heterogeneity).
  - We don't actually observe any one person's treatment effect ! (Missing Data problem)
- We need strong assumptions in order to recover $f(\beta_i)$ from data.
- Instead we can characterize simpler functions such as $E[\beta_i]$ (ATE) or $E[\beta_i|T_i = 1]$ (ATT) or $E[\beta_i|T_i = 0]$ (ATC) with fewer restrictions.

More Difficulties

What is hard here?

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

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

- Agents usually choose T_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.

Start with Easy Cases

- Let's start with the easy cases: run OLS and see what happens.
- OLS compares mean of treatment group with mean of control group (possibly controlling for other X)

$$\begin{aligned}\beta^{OLS} &= E(Y_i|T_i=1) - E(Y_i|T_i=0) \\ &= \underbrace{E[\beta_i|T_i=1]}_{\text{ATT}} + \left(\underbrace{E[u_i|T_i=1] - E[u_i|T_i=0]}_{\text{selection bias}}\right)\end{aligned}$$

- Even in absence of heterogeneity $\beta_i = \beta$ we can still have selection bias.
- $Y_i^0 = \alpha + u_i$ may vary within the population (this is quite common).

Solutions

1. Matching
2. Instrumental Variables
3. Diference 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.

Matching

- Compare treated individuals to un-treated individuals with identical observable characteristics X_i.
- Key assumption: everything about $Y_i(1) - Y_i(0)$ is captured in X_i; or u_i is randomly assigned conditional on X_i.
- Basic idea: The treatment group and the control group don't have the same distribution of observed characteristics as one another.
- Re-weight the un-treated population so that it resembles the treated population.
- Once distribution of X_i is the same for both groups $X_i|T_i \sim X_i$ then we assume all other differences are irrelevant and can just compare means.
- Matching assumes all selection is on observables.

Matching

- Formally the key assumption is the Conditional Independence Assumption (CIA)

$$\{Y_i^1, Y_i^0\} \perp T_i | X_i$$

- Once we know X_i allocation to treatment T_i is as if it is random.
- The only difference between treatment and control is composition of the sample.

Matching

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

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

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:

A Matching Example

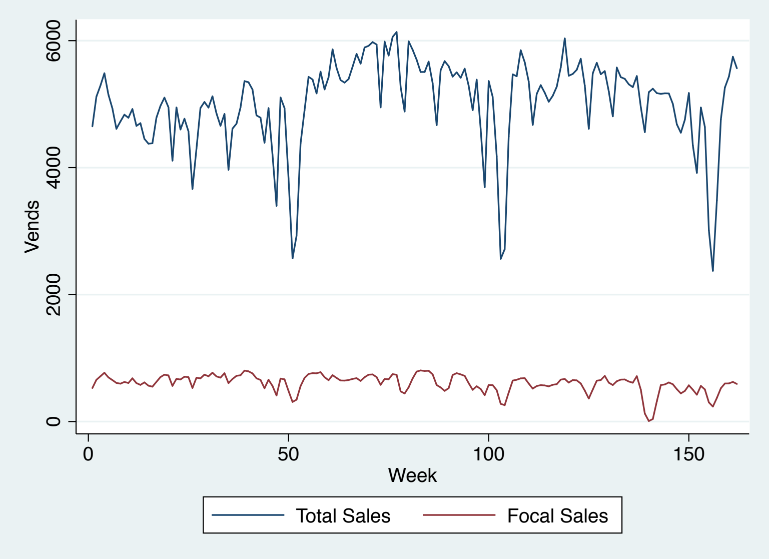
Here is an example where I found that matching was helpful in my own work with Julie Mortimer:

- 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!
- Can't use shares instead of quantities. Why?


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*	32.9
Twix Caramel	187.6	55.3	297.1*	100.0	109.5*	58.4
Assorted Chocolate	334.8	66.7	398.0*	95.0	63.2*	18.9
Assorted Energy	571.9	63.5	616.2	76.7	44.3	7.8
Zoo Animal Cracker	209.1	78.6	243.7*	98.1	34.6*	16.5
Salted Peanuts	187.9	70.4	216.3*	93.7	28.4	15.1
Choc Chip Famous Amos	171.6	71.7	193.1*	95.0	21.5*	12.5
Ruger Vanilla Wafer	107.3	59.7	127.9	78.6	20.6*	19.1
Assorted Candy	215.8	43.4	229.6	60.4	13.7	6.4
Assorted Potato Chips	279.6	64.2	292.4*	66.7	12.8	4.6
Assorted Pretzels	548.3	87.4	557.7*	88.7	9.4	1.7
Raisinets	133.3	66.0	139.4	74.2	6.1	4.6
Cheetos	262.2	60.1	260.5	58.2	-1.8	-0.7
Grandmas Choc Chip	77.9	51.3	72.5	37.8	-5.4	-7.0
Doritos	215.4	54.1	203.1	39.6	-12.3*	-5.7
Assorted Cookie	180.3	61.0	162.4	48.4	-17.9	-10.0
Skittles	100.1	62.9	75.1*	30.2	-25.1*	-25.0
Assorted Salty Snack	1382.8	56.0	1276.2*	23.3	-106.7*	-7.7
Snickers	323.4	50.3	2.0*	1.3	-321.4*	-99.4
Total	5849.6	74.2	5841.3	73.0	-8.3	-0.1

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.

Higher Dimensions

So matching works great in dimension 1. But what if $dim(X) > 1$?

- True high-dimensional matching may be infeasible. There may be no set of weights such that: $f(X_i|T_i = 1) \equiv \int w_i f(X_i|T_i = 0)\partial w_i$.
- One solution is the nearest-neighbor approach in Abadie Imbens (2006).
- 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

$$P(T_i = 1 | X_i) \equiv P(X_i)$$

- They prove that the propensity score and any function of X, $b(X)$ which is finer serves as a balancing score.
- Finer implies that:

$$\begin{aligned} b(X^1) = b(X^2) \quad &\Longrightarrow \quad P(X^1) = P(X^2) \\ P(X^1) = P(X^2) \quad &\not\Longrightarrow \quad b(X^1) = b(X^2) \end{aligned}$$

Propensity Score

- Main result: If treatment assignment is strongly ignorable conditional on X (CIA) then it is strongly ignorable $Y(1), Y(0) \perp T|X$ given any balancing score $b(X)$ including the propensity score:

$$\begin{aligned} Pr(T = 1|Y(1), Y(0), P(X)) &= E[Pr(T = 1|Y(1), Y(0), X)|P(X)] \\ &= E[Pr(T = 1|x)|P(X)] = P(X) \end{aligned}$$

- Also we require that $0 < P(X) < 1$ at each X which is known as the support condition.
- The theorem implies that given $P(X)$ we have as if random assignment.

Propensity Score

- Instead of matching on K dimensional X we can now match on a one-dimensional propensity score
- Thus the propensity score provides dimension reduction
- We still have to estimate the propensity score which is a high dimensional problem without *ad-hoc* parametric restrictions.
- Let us begin by assuming a can-opener.

Propensity Score

Just like in the matching case the problem arises because we do not observe the counterfactual mean:

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

With conditional independence and the propensity score:

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

Kernel Matching

How do we implement?

- Kernels are an obvious choice

$$\widehat{ATT} = \frac{1}{N_1} \sum_{i \in T=1} \left[Y_i - \frac{\sum_{j \in T=0} Y_j K\left(P(X_i) - P(X_j)\right)}{\sum_{s \in T=0} K\left(P(X_i) - P(X_s)\right)} \right]$$

  where N_1 is the sample size of the treatment group
  and $K(u)$ is a valid Kernel weight (people tend to use Gaussian Kernels here)

- As your propensity score gets further away from observation i you get less weight
- As $h \to 0$ (or σ_h) the window gets smaller and we use fewer neighbors.

Kernel Matching

- The usual caveats apply: h determines the bias-variance tradeoff
- Choice of Kernel effects finite-sample properties
- Here the common support is important. We can only learn about cases where $P(X) \neq 1$ and $P(X) \neq 0$. If you always get treated (or not-treated) we cannot learn from this observation.
- We also have to be careful in choosing X so as not to violate CIA (too many X's , too few X's) $\rightarrow$ have to think carefully!
- If you use propensity scores you will need a slide convincing us you have thought about why CIA holds for you!

Gotcha!

Under CIA we know

$$G(Y(1), Y(0)|X, T) = G(Y(1), Y(0)|X)$$

Suppose we add in Z, then we require that:

$$G(Y(1), Y(0)|X, Z, T) = G(Y(1), Y(0)|X, Z)$$

$$\begin{aligned} G(Y(1), Y(0)|X, T) &= \int G(Y(1), Y(0)|X, Z, T) dF(Z|X, T) \\ &= G(Y(1), Y(0)|X) \end{aligned}$$

where the last part follows by CIA.

- Thus each element can depend on T conditional on Z, X but the average may not.
- Mindless applications of matching can give you biased results!

Matching and OLS

- Recall that OLS is a special case of Kernel regression (and hence matching!)
- Think about

$$Y = \alpha + \beta T_i + u$$

- Assume that $E(u|T, X) = E(u|X)$ which is a conditional mean independence assumption
- The we can get β consistently (but not other variables) by running the following:

$$Y = \alpha + \beta T_i + \gamma X + v$$

- Again we are in the homogenous treatment world

What about IV

So what does IV do?

- Let's assume a binary instrument $Z_i = 1$
- $Y_i(1), Y_i(0)$ depends on the value of T_i
- But now we endogenize $T_i(1), T_i(0)$ where the argument is the value of Z_i.
- We observe $\{Z_i, T_i = T_i(Z_i), Y_i = Y_i(T_i(Z_i))\}$.

IV Assumptions

So what does IV do?

Independence $Z_i \perp Y_i(1), Y_i(0), T_i(1), T_i(0)$. Instrument is as if randomly assigned and does not directly affect Y_i

This is not implied by random assignment. In that case there would be four potential outcomes $Y_i(z, t)$

Random Assignment $Z_i \perp Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), T_i(1), T_i(0)$.

Exclusion Restriction $Y_i(z, t) = Y_i(z', t)$ for all z, z', t.

Thus we require both RA and ER to guarantee Independence. The second assumption is a substantive one.

We only observe (Z_i, T_i) not the pair $T_i(0), T_i(1)$ so we cannot determine compliance types directly! (See the picture)

IV Assumptions

Table 1: Compliance Types

		$W_i(0)$	
		0	1
$W_i(1)$	0	never-taker	defier
	1	complier	always-taker

IV Assumptions

We are stuck without further assumptions, so we assume:

Monotonicity/No Defiers $T_i(1) \geq T_i(0)$

- Works in many applications (classical drug compliance).
- Implied by many latent index models with constant coefficients
- Works as long as sign of $\pi_{1,i}$ doesn't change

$$T_i(z) = 1[\pi_0 + \pi_1 z + \varepsilon_i > 0]$$

IV Assumptions

Table 2: Compliance Type by Treatment and Instrument

		Z_i	
		0	1
W_i	0	complier/never-taker	never-taker/defier
	1	always-taker/defier	complier/always-taker

IV Assumptions

Table 3: Compliance Type by Treatment and Instrument given Monotonicity

		Z_i	
		0	1
W_i	0	complier/never-taker	never-taker
	1	always-taker	complier/always-taker

LATE Derivation

- We can derive the expression for β_{IV} as:

$$\beta_{IV} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[T_i|Z_i = 1] - E[T_i|Z_i = 0]} = E[Y_i(1) - Y_i(0)|complier]$$

- We can derive the expression for π_c (the fraction of compliers):

$$\pi_c = E[T_i|Z_i = 1] - E[T_i|Z_i = 0]$$

- Proof see Angrist and Imbens

How Close to ATE?

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

- $E[Y_i(0)|\text{never-taker}]$ and $E[Y_i(1)|\text{always-taker}]$ can be estimated from the data
- Compare these to their respective compliers $E[Y_i(0)|\text{complier}]$, $E[Y_i(1)|\text{complier}]$.
- When these are close then possibly $ATE \approx LATE$.

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
  - A J-statistic might reject even if both Z_1 and Z_2 are exogenous! (Why?).

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.

Local Average Treatment Effect

Finally, we need to understand what monotonicity means in terms of restrictions on economic theory.

- To quote from Vytlacil (2002) Econometrica:
  "The LATE assumptions are not weaker than the assumptions of a latent index model, but instead impose the same restrictions on the counterfactual data as the classical selection model if one does not impose parametric functional form or distributional assumptions on the latter."
- This is important because it shows that the LATE assumptions are equivalent to whatever economic modeling assumptions are required to justify the standard Heckman selection model and has no claim to greater generality.
- On the other hand there are no magical solutions to identifying effects when endogeneity/selection is present; this problem is exacerbated when the effects are heterogeneous and individuals select into treatment on the basis of the returns.

Further approaches to evaluation of program effects:

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:

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

Before and After

An even simpler estimator is the before and after or 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.

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

- 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).

Difference in Differences

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

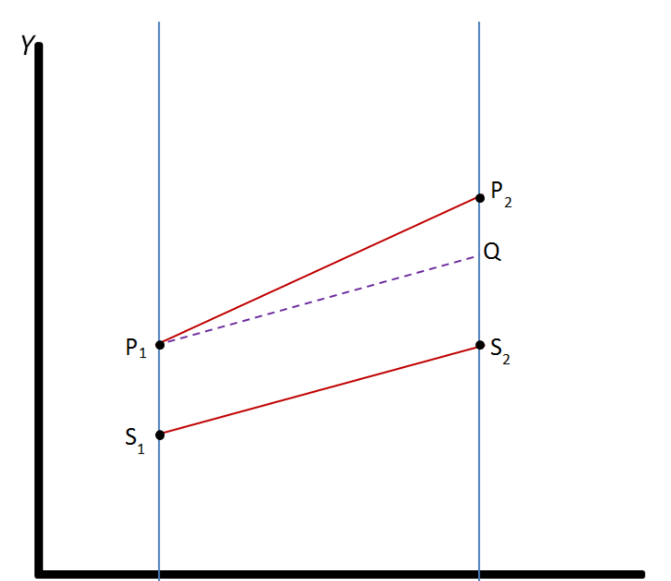
$$
\begin{aligned}
E[Y_{i2}^0 - Y_{i1}^0 | T_{i2} = 1] &= E[Y_{i2}^0 - Y_{i1}^0 | T_{i2} = 0] \\
E[Y_{i2} - Y_{i1} | T_{i2} = 0] &= d_2 - d_1
\end{aligned}
$$

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



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.

Difference in Differences

- #2 above violates the assumption $E[Y_{i2}^0 - Y_{i1}^0|T] = E[Y_{i2}^0 - Y_{i1}^0]$.
- To see why note that those participating into the program are such that $Y_{i0}^0 < B$. Assume for simplicity that the shocks u are iid. Hence $u_{i1} < B - \alpha_i - d_1$. This implies:

$$E[Y_{i2}^0 - Y_{i1}^0|T = 1] = d_2 = d_1 - E[u_{i1}|u_{i1} < B - \alpha_i - d_1]$$

For the control group:

$$E[Y_{i2}^0 - Y_{i1}^0|T = 1] = d_2 = d_1 - E[u_{i1}|u_{i1} > B - \alpha_i - d_1]$$

- Hence

$$E[Y_{i2}^0 - Y_{i1}^0|T = 1] - E[Y_{i2}^0 - Y_{i1}^0|T = 0] =$$
$$E[u_{i1}|u_{i1} > B - \alpha_i - d_1] - E[u_{i1}|u_{i1} < B - \alpha_i - d_1] > 0$$

- This is effectively regression to the mean: those unlucky enough to have a bad shock

Difference in Differences

Ashefelter (1978) was one of the first to consider difference in differences to evaluate training programs.

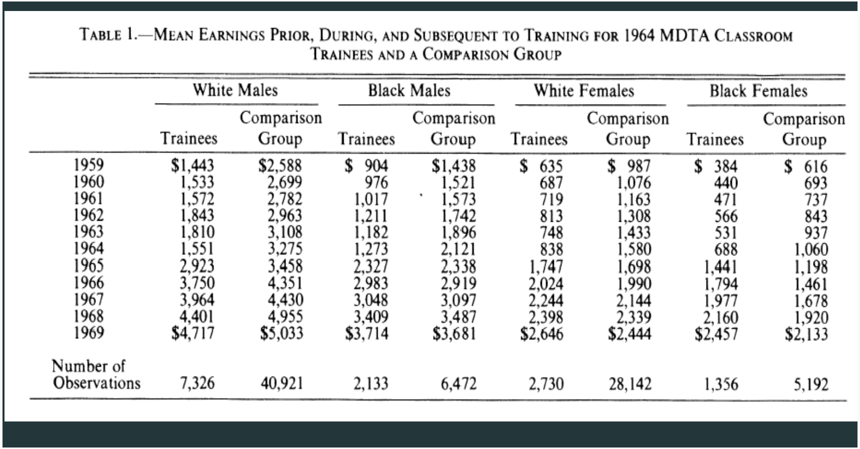


TABLE 1.—MEAN EARNINGS PRIOR, DURING, AND SUBSEQUENT TO TRAINING FOR 1964 MDTA CLASSROOM TRAINEES AND A COMPARISON GROUP

	White Males		Black Males		White Females		Black Females	
	Trainees	Comparison Group	Trainees	Comparison Group	Trainees	Comparison Group	Trainees	Comparison Group
1959	$1,443	$2,588	$ 904	$1,438	$ 635	$ 987	$ 384	$ 616
1960	1,533	2,699	976	1,521	687	1,076	440	693
1961	1,572	2,782	1,017	1,573	719	1,163	471	737
1962	1,843	2,963	1,211	1,742	813	1,308	566	843
1963	1,810	3,108	1,182	1,896	748	1,433	531	937
1964	1,551	3,275	1,273	2,121	838	1,580	688	1,060
1965	2,923	3,458	2,327	2,338	1,747	1,698	1,441	1,198
1966	3,750	4,351	2,983	2,919	2,024	1,990	1,794	1,461
1967	3,964	4,430	3,048	3,097	2,244	2,144	1,977	1,678
1968	4,401	4,955	3,409	3,487	2,398	2,339	2,160	1,920
1969	$4,717	$5,033	$3,714	$3,681	$2,646	$2,444	$2,457	$2,133
Number of Observations	7,326	40,921	2,133	6,472	2,730	28,142	1,356	5,192

Difference in Differences

Ashenfelter (1978) reports the following results.

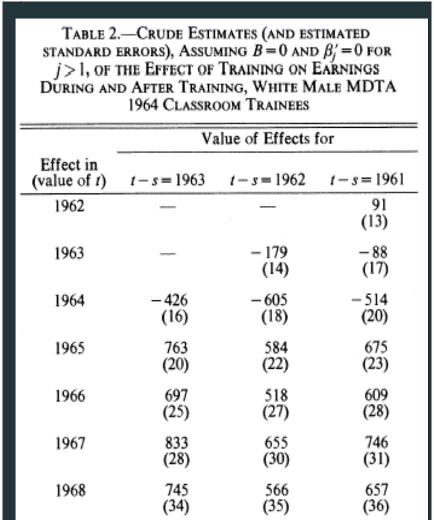


Table 2.—Crude Estimates (and Estimated Standard Errors), Assuming $B=0$ and $\beta_j'=0$ for $j>1$, of the Effect of Training on Earnings During and After Training, White Male MDTA 1964 Classroom Trainees

Effect in (value of t)	Value of Effects for $t-s=1963$	 $t-s=1962$	 $t-s=1961$
1962	—	—	91 (13)
1963	—	−179 (14)	−88 (17)
1964	−426 (16)	−605 (18)	−514 (20)
1965	763 (20)	584 (22)	675 (23)
1966	697 (25)	518 (27)	609 (28)
1967	833 (28)	655 (30)	746 (31)
1968	745 (34)	566 (35)	657 (36)

Difference in Differences

- 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.

Difference in Differences 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. We can still use difference in differences.
- First consider a simple case where $E[Y_{i2}^0 - Y_{i1}^0|T] = E[Y_{i2}^0 - Y_{i1}^0]$.
- We need to modify slightly the assumption to

$$\begin{aligned} E[Y_{i2}^0|\text{Group receiving training}] - E[Y_{i1}^0|\text{Group receiving training in the next period}] \\ = E[Y_{i2}^0 - Y_{i1}^0] \end{aligned}$$

  which requires, in addition to the original independence assumption that conditioned on particular individuals that population we will be sampling from does not change composition.
- We can then obtain immediately an estimator for ATT as

$$E[\beta_i|T_{i2} = 1]$$

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).

Difference in Differences and Selection on Unobservables

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

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

  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

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

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

RDD

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 this is taken from the JEL Paper by Lee and Lemieux (2010).

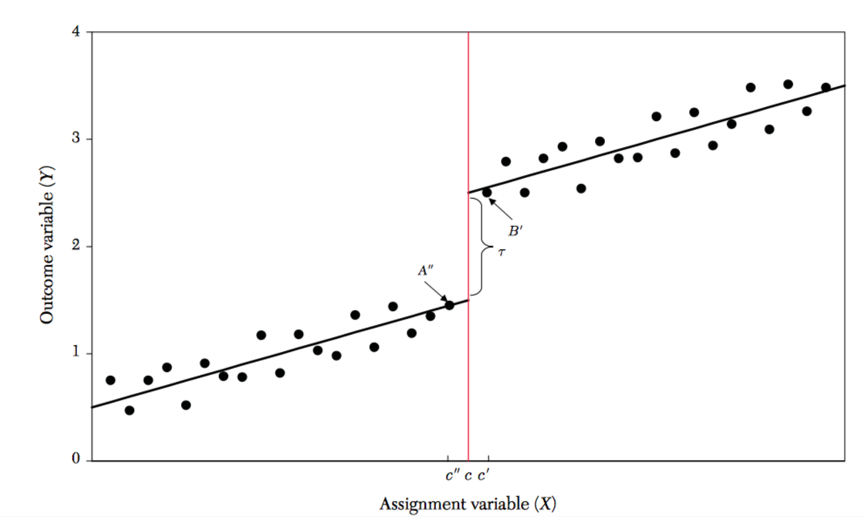
RDD: Basics

- 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 → draw pictures!
- We could run the regression where $D_i = \mathbf{1}[X_i > c]$.

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

- This puts a lot of restrictions (linearity) on the relationship between Y and X.
- Also (without additional assumptions) we only learn about τ_i at the point $X = c$.

RDD: Nonlinearity

First thing to relax is assumption of linearity.

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

This is known as partially linear model.

- 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

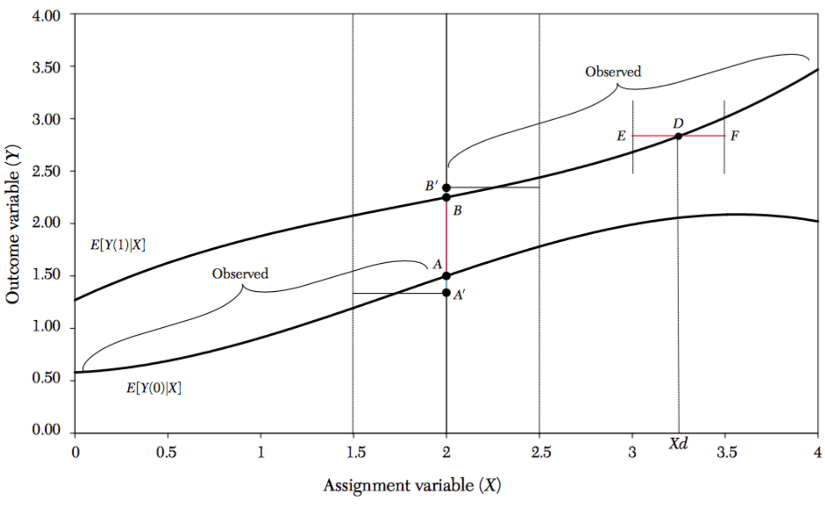


Figure 2. Nonlinear RD

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: Lee (2008)

Looked at incumbency advantage in the US House of Representatives

- Running variable was vote share in previous election
    - Problem of naive approach: good candidates get lots of votes!
    - Compare outcomes of districts with barely D to barely R.
- First we plot bin-scatter plots and quartic (from each side) polynomials.
- Discussion about how to choose bin-scatter bandwidth (CV).

Lee (2008)

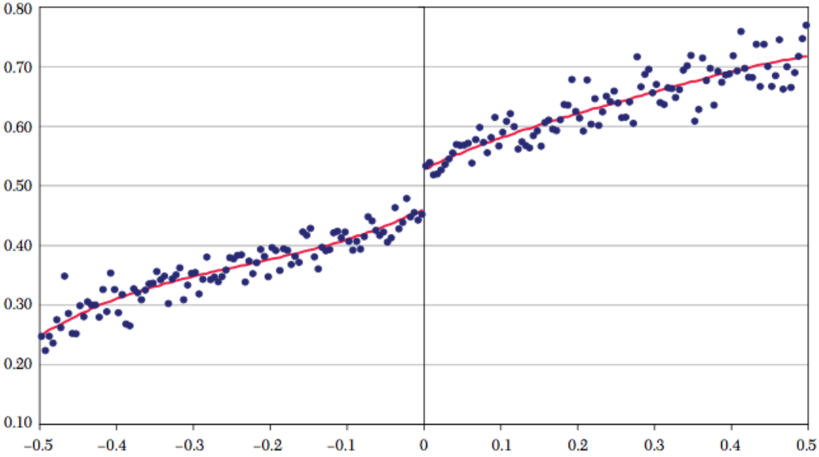


Figure 8. Share of Vote in Next Election, Bandwidth of 0.005 (200 bins)

Lee (2008)

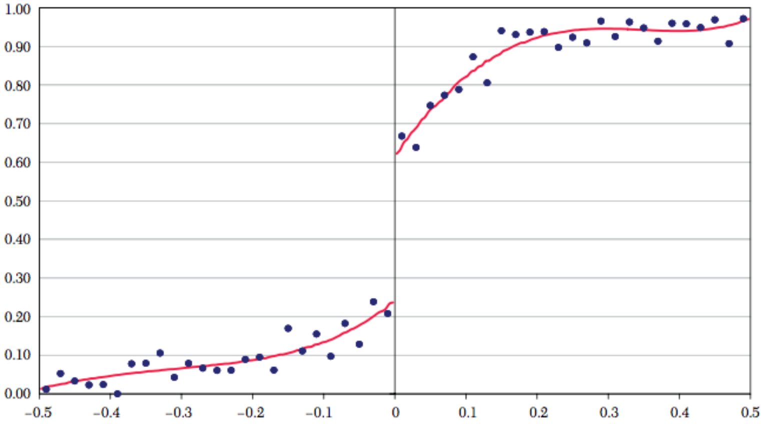


Figure 9. Winning the Next Election, Bandwidth of 0.02 (50 bins)

Lee (2008)

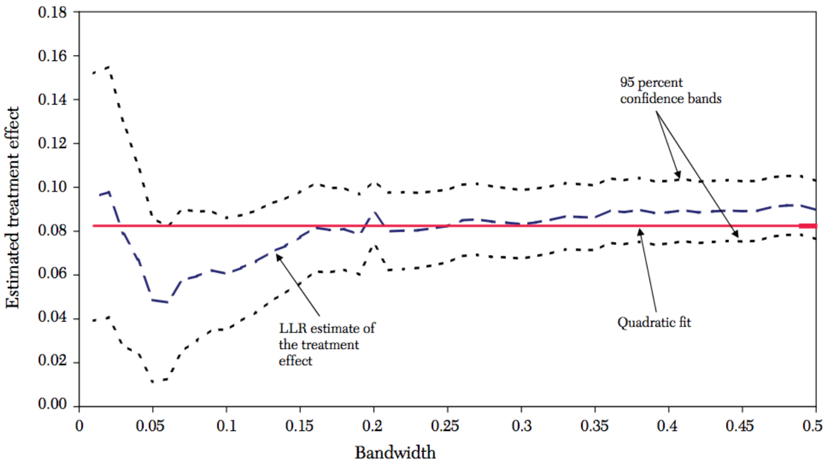


Figure 18. Local Linear Regression with Varying Bandwidth: Share of Vote at Next Election

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 $x = c$.

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.

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 $U_1 - U_0 \equiv \eta$.

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

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

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).

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

- 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

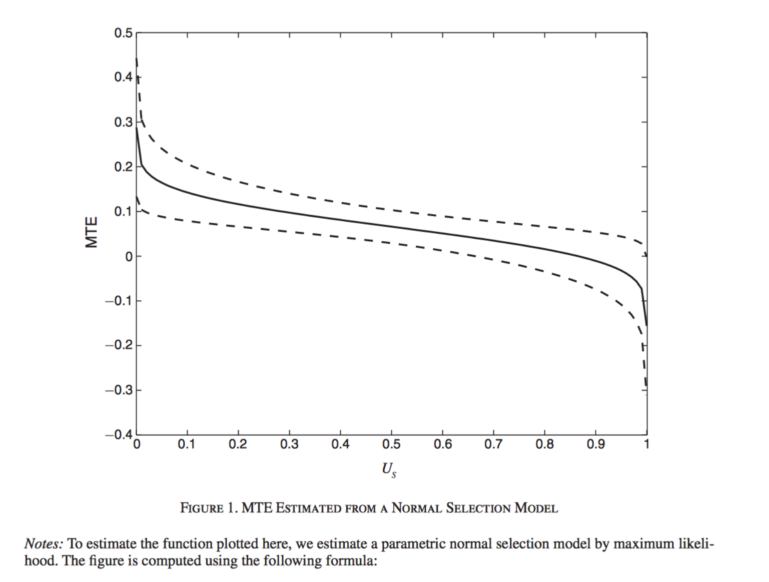


FIGURE 1. MTE ESTIMATED FROM A NORMAL SELECTION MODEL

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:

Carneiro, Heckman and Vytlacil

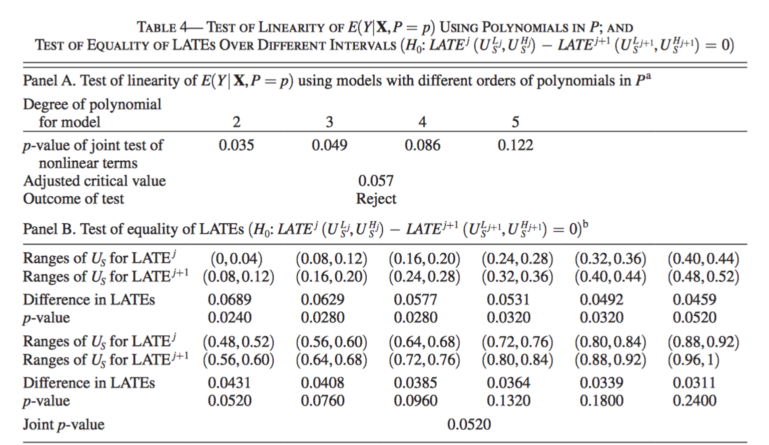


TABLE 4—TEST OF LINEARITY OF $E(Y|\mathbf{X}, P = p)$ USING POLYNOMIALS IN P; AND TEST OF EQUALITY OF LATEs OVER DIFFERENT INTERVALS (H_0: $LATE^j\,(U_S^{Lj}, U_S^{Hj}) - LATE^{j+1}\,(U_S^{Lj+1}, U_S^{Hj+1}) = 0$)

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

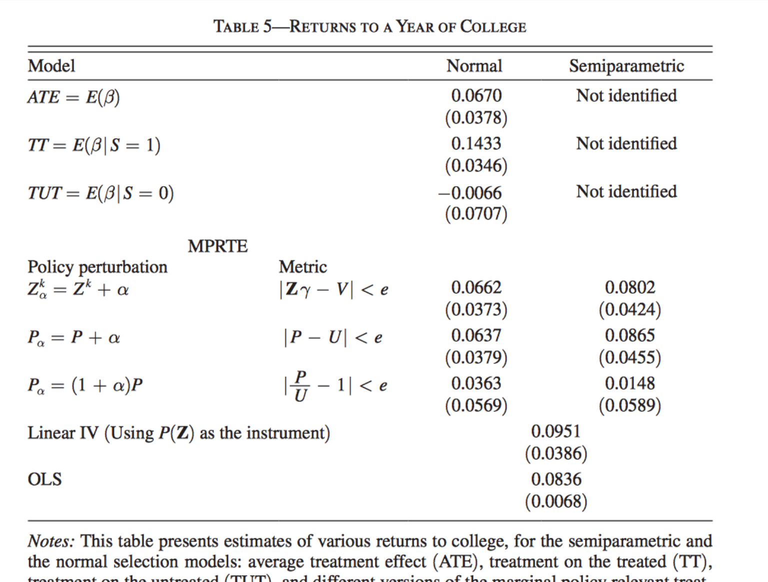


TABLE 5—RETURNS TO A YEAR OF COLLEGE

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 treat

Carneiro, Heckman and Vytlacil

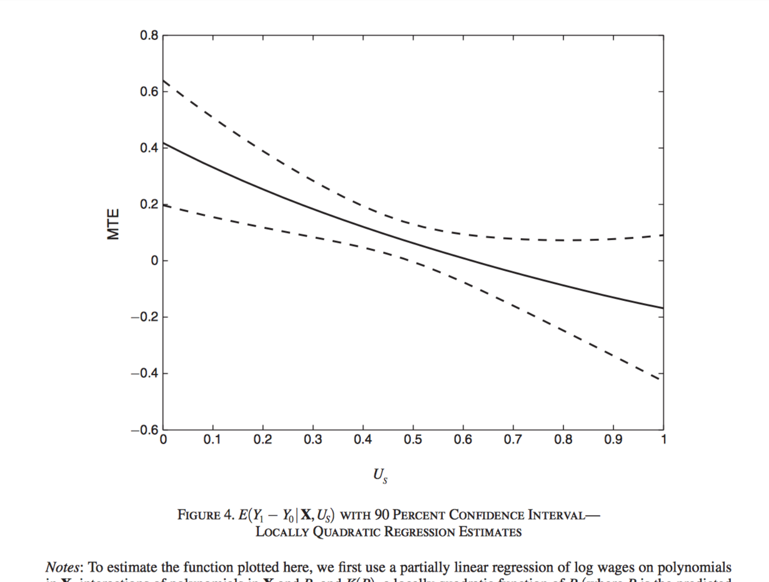


FIGURE 4. $E(Y_1 - Y_0 | \mathbf{X}, U_S)$ WITH 90 PERCENT CONFIDENCE INTERVAL—
LOCALLY QUADRATIC REGRESSION ESTIMATES

Notes: To estimate the function plotted here, we first use a partially linear regression of log wages on polynomials

Carneiro, Heckman and Vytlacil

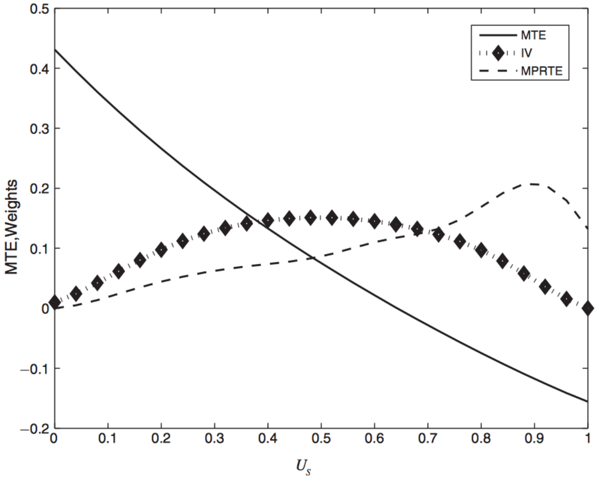


FIGURE 6. WEIGHTS FOR IV AND MPRTE

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

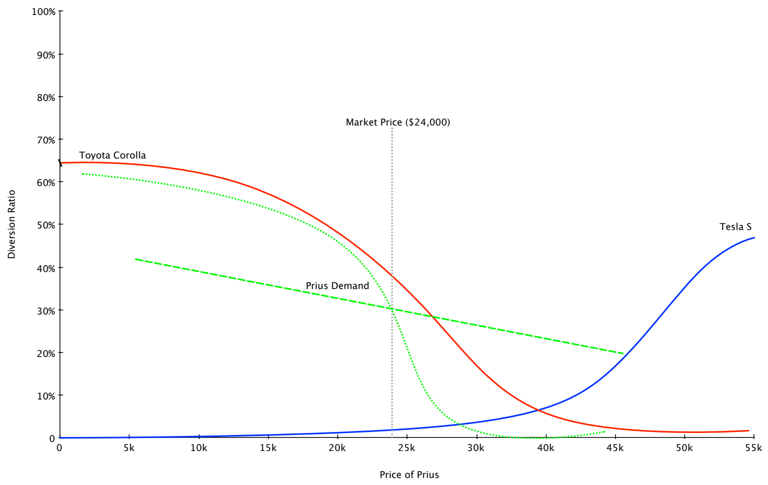
I have done some work trying to bring these methods into merger analysis.

- 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!