14.385 Nonlinear Econometric Analysis

Conditional independence, Difference in Differences, Regression Discontinuity

Maximilian Kasy

Department of Economics, MIT

Fall 2022

Takeaways for this part of class

1. fundamental notions of causal inference:
   - causality
   - structural objects
   - identification
2. identification approaches:
   - randomized experiments
   - instrumental variables
   - conditional independence
   - difference in differences
   - regression discontinuity
3. analog estimators

Roadmap

Conditional independence

Difference in Differences

Regression Discontinuity

Analog estimators

Conditional independence

- Causal identification approaches as generalizations of randomized experiments!
- last section:
  1. Generalizing from $\boldsymbol{D}$ random to $\boldsymbol{Z}$ random.
- this section:
  Generalizing from $\boldsymbol{D}$ random
  to $\boldsymbol{D}$ random conditional on covariates $\boldsymbol{X}$.
- If that is the case,
  "comparing apples with apples"
  only requires comparing units with the same values of $\boldsymbol{X}$.

Assumptions

1. $D \in \{0,1\}$
2. $Y = D \cdot Y^1 + (1-D) \cdot Y^0$
3. $(Y^1, Y^0) \perp D|X$
4. $0 < p(X) < 1$ for almost all X, where $p(X) = P(D = 1|X)$

Discussion of assumptions

1. **Binary treatment:**
   easy to generalize the following to arbitrary support of D.
2. **Potential outcome equation for Y:** $Y = D \cdot Y^1 + (1 - D) \cdot Y^0$
   same as before (SUTVA!).
3. **Conditional independence:** $(Y^1, Y^0) \perp D | X$
   - within subgroups defined by X, we essentially have a randomized experiment.
   - somewhat hard to justify in practice
   - best thought of as a plausible approximation, if X captures the main components of heterogeneity which might drive dependence between potential outcomes and treatment

4. **Overlapping support:** $0 < p(X) < 1$
   - there are no groups, defined by X, for which everybody (or nobody) was treated.
   - can be checked directly in the data (like instrument relevance)

Under these assumptions, the average treatment effect is identified

Practice problem

Try to prove this!

Identification of ATE, proof using regression

- conditional expectation of outcomes given covariates and treatment:

$$E[Y|X, D = 1] = E[Y^1|X, D = 1] = E[Y^1|X].$$

- first equality holds by the potential outcomes equation.
- second equality uses conditional independence.
- $E[Y|X, D = 1]$ is identified as long as $p(X) > 0$.
- law of iterated expectations,
  averaging across the distribution of $X \Rightarrow$

$$E[Y^1] = E_X[E[Y^1|X]] = E_X[E[Y|X, D = 1]].$$

- note that
$$E_X[E[Y|X, D = 1]] \neq E[Y|D = 1]!$$
- left hand term averages $E[Y|X, D = 1]$ over the marginal distribution of X
- right hand term averages $E[Y|X, D = 1]$ over the conditional distribution of X given D.
- similarly for $D = 0$
$$E[Y^0] = E_X[E[Y^0|X]] = E_X[E[Y|X, D = 0]].$$
- average treatment effect is identified by
$$E[Y^1 - Y^0] = E_X[E[Y|X, D = 1] - E[Y|X, D = 0]].$$

Identification of ATE, proof using reweighting

- intuition: units for which $p(X)$ is small
  are under-represented among the observations such that $D = 1$
- need to upweight them in order to get the distribution of Y^1
- upweight by factor $1/p(x)$ – "inverse probability weighting"

- Law of iterated expectations, D binary $\Rightarrow$

$$\begin{aligned} E[YD|X] &= E[YD|X, D=1] \cdot p(X) + E[YD|X, D=0] \cdot (1-p(X)) \\ &= E[Y|X, D=1] \cdot p(X) \end{aligned}$$

- Potential outcome equation, conditional independence $\Rightarrow$

$$E[Y|X, D=1] = E[Y^1|X]$$

- Rearranging $\Rightarrow$

$$E\left[\frac{D}{p(X)} \cdot Y \middle| X\right] = E[Y^1|X]$$

- Iterated expectations, again $\Rightarrow$

$$E\left[\frac{D}{p(X)} \cdot Y\right] = E[Y^1]$$

- Similar argument for $D = 0 \Rightarrow$

$$E\left[\frac{1-D}{1-p(X)} \cdot Y\right] = E[Y^0]$$

Practice problem

Show this.

- $\Rightarrow$ Average treatment effect is identified by

$$\begin{aligned} E[Y^1 - Y^0] &= E\left[\left(\frac{D}{p(X)} - \frac{1-D}{1-p(X)}\right) \cdot Y\right] \\ &= E\left[\left(\frac{D-p(X)}{p(X)(1-p(X)}\right) \cdot Y\right]. \end{aligned}$$

Difference in Differences

- Causal identification approaches as generalizations of randomized experiments!
- Previous generalizations of randomized experiments:
  1. Generalizing from D random to Z random.
  2. Generalizing from D random to D random conditional on covariates X.
- This section: Generalizing from D random to $D_1 - D_0$ (change over time) random relative to $Y_1^0 - Y_0^0$ (counterfactual trend).
- "Common trends"

Difference-in-Differences Setup

- Two groups:
  - $D = 1$: treated units
  - $D = 0$: control units
- Two periods:
  - $t = 0$: pre-treatment period
  - $t = 1$: post-treatment period
- Potential outcomes:
  - Y_t^1: outcome in period t if treated before t
  - Y_t^0: outcome in period t if not treated before t

Difference-in-Differences Setup

- Treatment effect for unit i at time t is

$$Y_t^1 - Y_t^0.$$

- Observed outcomes Y_t are realized as

$$Y_t = Y_t^0(1 - D_t) + Y_t^1 D_t.$$

- Because the treatment occurs only after $t = 0$, we define

$$D = D_1.$$

- It follows that,

$$\begin{aligned} Y_0 &= Y_0^0, \\ Y_1 &= Y_1^0(1 - D) + Y_1^1 D. \end{aligned}$$

- Consider the average treatment effect on the treated,

$$ATT = E[Y_1^1 - Y_1^0 | D = 1].$$

- **Common trends assumption:**

$$E[Y_1^0 - Y_0^0 | D = 1] = E[Y_1^0 - Y_0^0 | D = 0],$$

  that is, the treated and non-treated would have exhibited the same trend in the absence of the treatment.
- Under this assumption, the *ATT* is identified.

Practice problem

Prove this.

Identification of ATT under common trends, proof

- By common trends,

$$
\begin{aligned}
&(E[Y_1|D=1] - E[Y_1|D=0]) - (E[Y_0|D=1] - E[Y_0|D=0]) \\
&= \Big(E[Y_1^1|D=1] - E[Y_1^0|D=0]\Big) - \Big(E[Y_0^0|D=1] - E[Y_0^0|D=0]\Big) \\
&= \Big(E[Y_1^1|D=1] - E[Y_1^0|D=1]\Big) = ATT.
\end{aligned}
$$

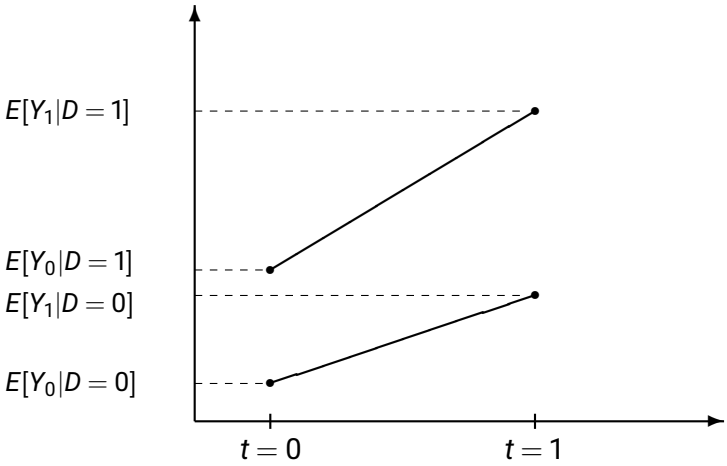
Practice problem

Suppose you are interested in the ATT of D on $\log Y$,

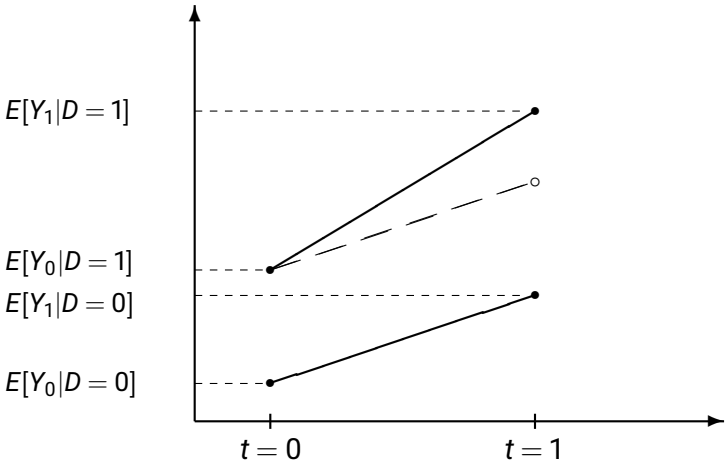
$$ATT = E[\log Y_1^1 - \log Y_1^0 | D=1].$$

Is this effect identified under the above assumptions?

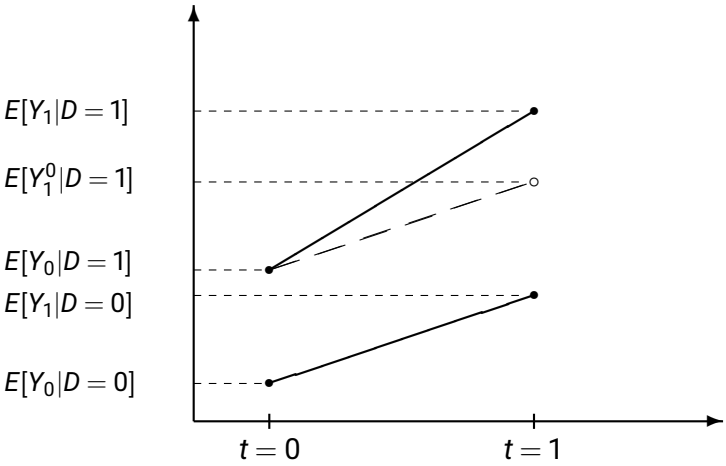
Difference-in-Differences: Graphical Interpretation



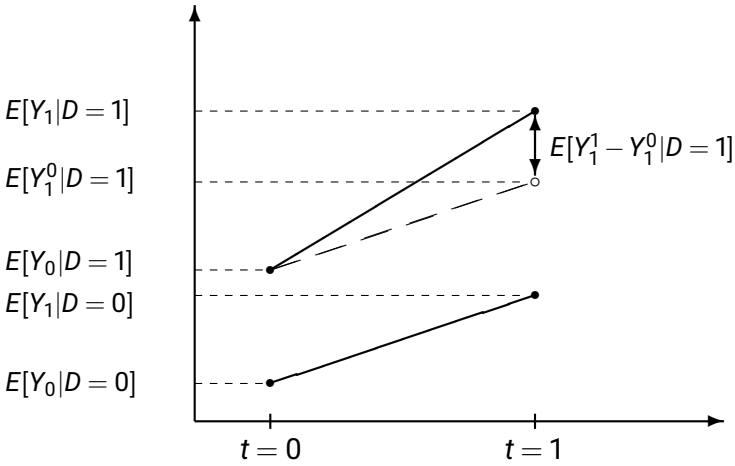
Difference-in-Differences: Graphical Interpretation



Difference-in-Differences: Graphical Interpretation



Difference-in-Differences: Graphical Interpretation



Empirical example

- **Card, D. (1990).** The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review*, 43(2):245–257.
- The Mariel Boatlift from Cuba in 1980 increased the Miami labor force by 7%
- Comparing individual-level data on unemployment from the Current Population Survey (CPS) for Miami and four comparison cities (Atlanta, Los Angeles, Houston and Tampa-St. Petersburg)

Table 4
Differences-in-differences estimates of the effect of immigration on unemployment[a]

	Group	Year		
		1979 (1)	1981 (2)	1981–1979 (3)
	Whites			
(1)	Miami	5.1 (1.1)	3.9 (0.9)	−1.2 (1.4)
(2)	Comparison cities	4.4 (0.3)	4.3 (0.3)	−0.1 (0.4)
(3)	Miami-Comparison Difference	0.7 (1.1)	−0.4 (0.95)	−1.1 (1.5)
	Blacks			
(4)	Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
(5)	Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
(6)	Miami-Comparison Difference	−2.0 (1.9)	−3.0 (2.0)	−1.0 (2.8)

[a] Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

Empirical example

- **Qian, N. (2008).** Missing women and the price of tea in China: The effect of sex-specific earnings on sex imbalance. *The Quarterly Journal of Economics*, 123(3):1251–1285.
- Traditional tea growing requires more female labor, orchards more male labor.
- These crops can only be grown in some regions.
- Post 1979 reforms increased the value of these crops.
- Finding: Increasing female income, holding male income constant, improves survival rates for girls, whereas increasing male income, holding female income constant, worsens survival rates for girls.

TABLE III

OLS AND 2SLS ESTIMATES OF THE EFFECT OF PLANTING TEA AND ORCHARDS ON SEX RATIOS CONTROLLING FOR COUNTY LEVEL LINEAR COHORT TRENDS

	Dependent variables					
	Fraction of males			Tea × post	Fraction of males	
	(1) OLS	(2) OLS	(3) OLS	(4) 1st	(5) IV	(6) IV
Tea × post	−0.012	−0.013	−0.012		−0.072	−0.011
	(0.007)	(0.006)	(0.005)		(0.031)	(0.007)
Orchard × post	0.005					
	(0.002)					
Slope × post	−0.002			0.26		
	(0.002)			(0.057)		
Linear trend	No	No	Yes	Yes	No	Yes
Observations	28,349	37,756	37,756	37,756	37,756	37,756

Notes. Coefficients of the interactions between dummies indicating whether a cohort was born post-reform

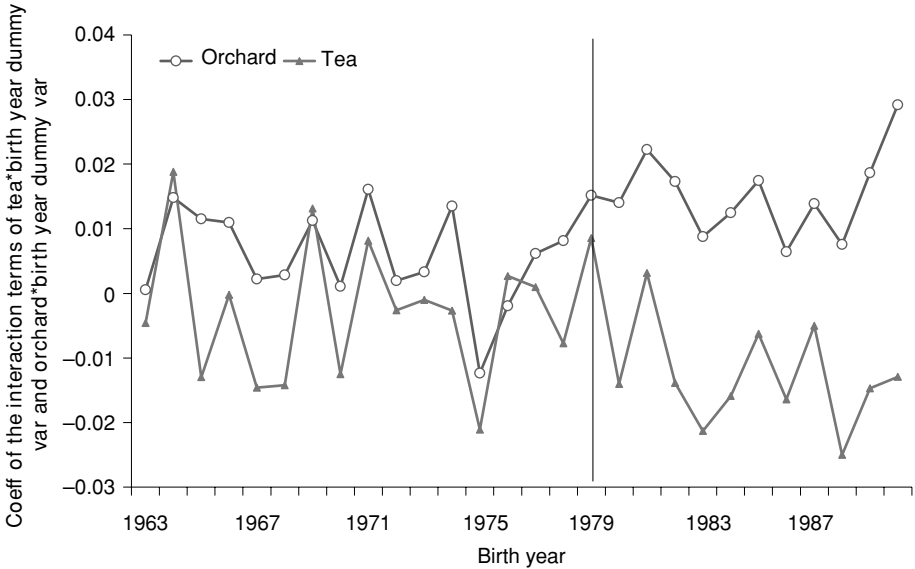


FIGURE V
The Effect of Planting Tea and Orchards on Sex Ratios

Regression Discontinuity

- Causal identification approaches as generalizations of randomized experiments!
- Previous generalizations of randomized experiments:
  1. Generalizing from D random to Z random.
  2. Generalizing from D random to D random conditional on covariates X.
  3. Generalizing from D random to $(D_1 - D_0)$ random relative to $(Y_1^0 - Y_0^0)$
- This section: Generalizing from D random,
  which implies $E[Y^1|D = d]$ and $E[Y^0|D = d]$ constant in d,
  to $E[Y^1|X = x]$ and $E[Y^0|X = x]$ continuous in x.

Sharp Regression Discontinuity Design

- **Discontinuous assignment of treatment**: Suppose assignment for treatment D is determined based on whether a unit exceeds some threshold c on a running variable X,

$$D = \mathbf{1}(X \geq c) = \begin{cases} 1 & \text{if } X \geq c \\ 0 & \text{if } X < c. \end{cases}$$

- **Continuous expectation of potential outcomes**: $E[Y^1|X = x]$ and $E[Y^0|X = x]$ are continuous in x.
- Under these assumptions, the conditional average treatment effect given $X = c$, $E[Y^1 - Y^0|X = c]$, is identified.

Practice problem

Prove this.

Identification for sharp RDD, proof

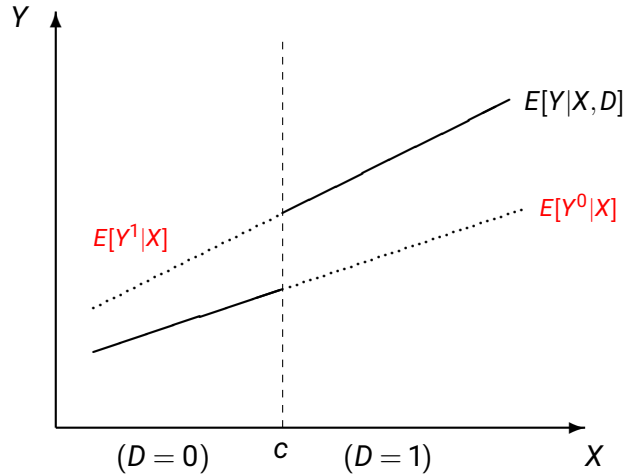
- Consider the right and left hand limits,

$$\begin{aligned}\lim_{x \downarrow c} E[Y|X=x] &= \lim_{x \downarrow c} E[Y^1|X=x] \\ &= E[Y^1|X=c] \\ \lim_{x \uparrow c} E[Y|X=x] &= \lim_{x \uparrow c} E[Y^0|X=x] \\ &= E[Y^0|X=c].\end{aligned}$$

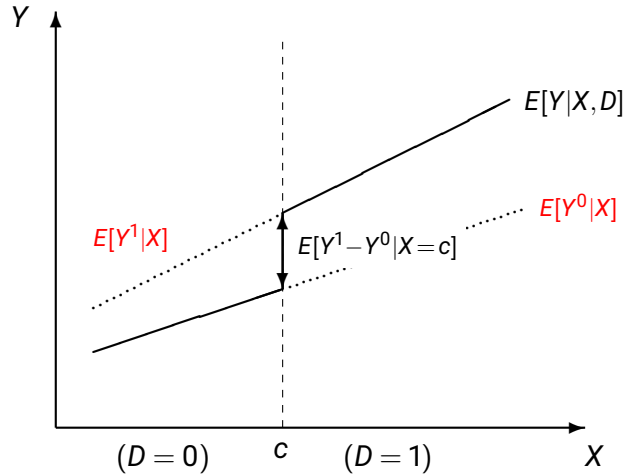
Remarks

- Design arises often from administrative or legal rules.
- Usually X is correlated with the potential outcomes Y^1, Y^0, so comparing treated and untreated does not provide causal estimates.
- But we can use the discontinuity in $E[Y|X]$ at the cutoff value $X = c$ to estimate the effect of D on Y for units with $X = c$.
- RDD is a fairly old idea (Thistlethwaite and Campbell, 1960) but this design experienced a renaissance in recent years.
- E.g., scholarships are given on the basis of whether or not the student's test score is larger than some cutting value.
  - Treatment D is scholarship
  - Forcing variable X is SAT score with cutoff c
  - Outcome Y is subsequent college grades
  - Y^0 denotes potential grades without the scholarship
  - Y^1 is potential grades with the scholarship

Sharp RDD: Graphical Interpretation



Sharp RDD: Graphical Interpretation



Empirical example

- **Lee (2001).** The Electoral Advantage to Incumbency and Voters' Valuation of Politicians' Experience: A Regression Discontinuity Analysis of Elections to the U.S. House *NBER working paper*.
- Incumbent parties and candidates enjoy great electoral success in the U.S. and other countries
- Measuring incumbent advantage is difficult because "better" parties or candidates may be consistently favored by the electorate
- But close elections might be "almost" random.
- Compare elections to the U.S. House of Representatives (1946 to 1998), where Democrats won but almost lost, versus elections where they lost but almost won.

Figure IVa: Democrat Party's Vote Share in Election t+1, by Margin of Victory in Election t: local averages and parametric fit

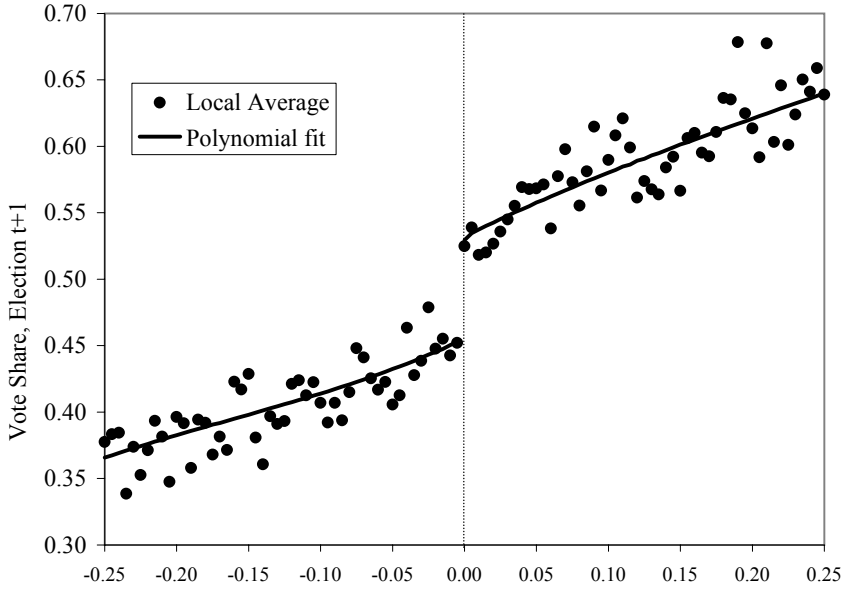
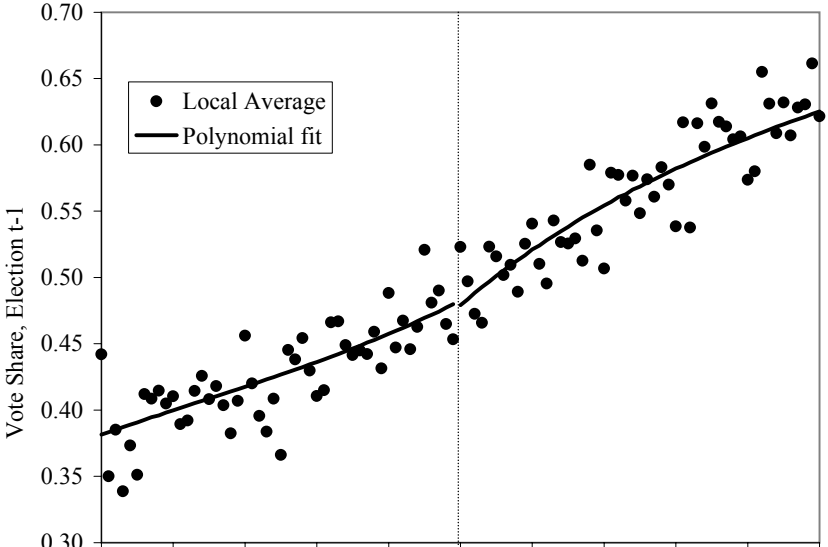


Figure IVb: Democratic Party Vote Share in Election t-1, by Margin of Victory in Election t: local averages and parametric fit



Fuzzy Regression Discontinuity Design

- Often: Cutoff does not perfectly determine treatment but creates a discontinuity in the probability of receiving the treatment
- For example: The probability of being offered a scholarship may jump at a certain SAT score, above which the applications are given "special consideration."
- **Fuzzy RD assumptions**:

$$
\begin{aligned}
Z &= \mathbf{1}(X > c) \\
D &= Z \cdot D^1 + (1 - Z) \cdot D^0 \\
Y &= D \cdot Y^1 + (1 - D) \cdot Y^0 \\
D^1 &\geq D^0,
\end{aligned}
$$

  and $E[(Y^1 D^1, Y^1 D^0, Y^0(1 - D^1), Y^0(1 - D^0), D^1, D^0) | X = x]$ is continuous in x.
- This design combines features of sharp RD and of IV.

Identification for fuzzy RDD

- Under these assumption, the conditional local average treatment effect

$$E[Y^1 - Y^0 | X = c, D^1 > D^0]$$

  is identified.

Practice problem

Prove this.

Proof

- Consider the right and left hand limits,

$$\begin{aligned}
\lim_{x \downarrow c} E[D|X=x] &= \lim_{x \downarrow c} E[D^1|X=x] = E[D^1|X=c] \\
\lim_{x \uparrow c} E[D|X=x] &= \lim_{x \uparrow c} E[D^0|X=x] = E[D^0|X=c] \\
\lim_{x \downarrow c} E[Y|X=x] &= \lim_{x \downarrow c} E[D^1 \cdot Y^1 + (1-D^1) \cdot Y^0|X=x] \\
&= E[D^1 \cdot Y^1 + (1-D^1)|X=c] \\
\lim_{x \uparrow c} E[Y|X=x] &= \lim_{x \uparrow c} E[D^0 \cdot Y^1 + (1-D^0) \cdot Y^0|X=x] \\
&= E[D^0 \cdot Y^1 + (1-D^0) \cdot Y^0|X=c].
\end{aligned}$$

Proof continued

- Thus

$$\lim_{x \downarrow c} E[Y|X = x] - \lim_{x \uparrow c} E[Y|X = x] = E[(Y^1 - Y^0) \cdot (D^1 - D^0)|X = c]$$
$$\lim_{x \downarrow c} E[D|X = x] - \lim_{x \uparrow c} E[D|X = x] = E[(D^1 - D^0)|X = c]$$

- and therefore

$$E[Y^1 - Y^0|X = c, D^1 > D^0] = \frac{\lim_{x \downarrow c} E[Y|X = x] - \lim_{x \uparrow c} E[Y|X = x]}{\lim_{x \downarrow c} E[D|X = x] - \lim_{x \uparrow c} E[D|X = x]}.$$

Recap of identification results

Throughout: $Y = D \cdot Y^1 + (1-D) \cdot Y^0$.

1. **Randomized experiments**: If $(Y^0, Y^1) \perp D$ then

$$E[Y|D=1] - E[Y|D=0] = E[Y^1 - Y^0] = ATE.$$

2. **Instrumental variables**: If
   - $D = Z \cdot D^1 + (1-Z) \cdot D^0$
   - $D^1 \geq D^0$
   - $Z \perp (Y^0, Y^1, D^0, D^1)$
   - $\text{Cov}(Z, D) \neq 0$

   then

$$\begin{aligned} \frac{\text{Cov}(Z,Y)}{\text{Cov}(Z,D)} &= \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} \\ &= E[Y^1 - Y^0 | D^1 > D^0] = LATE. \end{aligned}$$

3. **Conditional independence**: If $(Y^1, Y^0) \perp D|X$ then

$$E_X[E[Y|X,D=1]-E[Y|X,D=0]] = E\left[\left(\frac{D-p(X)}{p(X)(1-p(X))}\right)\cdot Y\right]$$
$$= E[Y^1 - Y^0] = ATE.$$

4. **Difference in differences**: If $E[Y_1^0 - Y_0^0|D=1] = E[Y_1^0 - Y_0^0|D=0]$ then:

$$(E[Y_1|D=1]-E[Y_1|D=0]) - (E[Y_0|D=1]-E[Y_0|D=0])$$
$$= E[Y_1^1 - Y_1^0|D=1] = ATT.$$

5. **Sharp regression discontinuity**: If $E[Y^1|X=x]$ and $E[Y^0|X=x]$ continuous in x at $x=c$, $D=\mathbf{1}(X>c)$, then

$$\lim_{x\downarrow c} E[Y|X=x] - \lim_{x\uparrow c} E[Y|X=x] = E[Y^1 - Y^0|X=c].$$

Brief remark on estimation

- This entire part of class was about **identification**.
- Mapping features of the population distribution into features of the underlying structure.
- Suggests analog-approaches for estimation:
- Replace expectations by sample means, e.g.

$$E[Y] \to \bar{Y} = \frac{1}{n}\sum_i Y_i$$

- Replace conditional expectations by sub-sample means, e.g.

$$E[Y|D=1] \to \bar{Y}|_{D=1} = \sum_i D_i Y_i \Big/ \sum_i D_i,$$

or predicted values of regressions, e.g.

$$E[Y|D,X] \to \alpha D + \beta X + \gamma D \cdot X.$$

Analog estimators

1. **Randomized experiments**:

$$\widehat{ATE} = \bar{Y}|_{D=1} - \bar{Y}|_{D=0}.$$

2. **Instrumental variables**:

$$\widehat{LATE} = \frac{\widehat{\mathrm{Cov}}(Z,Y)}{\widehat{\mathrm{Cov}}(Z,D)} = \frac{\bar{Y}|_{Z=1} - \bar{Y}|_{Z=0}}{\bar{D}|_{Z=1} - \bar{D}|_{Z=0}}.$$

3. **Conditional independence**: If X is discrete,

$$\widehat{E}[Y|X,D] = \bar{Y}|_{X,D}, \qquad \widehat{p}(X) = \bar{D}|_{X=x},$$

$$\begin{aligned}\widehat{ATE} &= \frac{1}{n}\sum_i \left(\bar{Y}|_{X=X_i,D=1} - \bar{Y}|_{X=X_i,D=0}\right) \\ &= \frac{1}{n}\sum_i \left(\frac{D_i - \bar{D}|_{X=X_i}}{\bar{D}|_{X=X_i}(1-\bar{D}|_{X=X_i})}\right)\cdot Y_i\end{aligned}$$

4. **Difference in differences**:

$$\widehat{ATT} = \left(\bar{Y}_1|_{D=1} - \bar{Y}_1|_{D=0}\right) - \left(\bar{Y}_0|_{D=1} - \bar{Y}_0|_{D=0}\right)$$

5. **Sharp regression discontinuity**:
   - Idea 1: replace limit by local average,

$$\widehat{\lim_{x \downarrow c} E}[Y|X = x] = \bar{Y}|_{c<X<c+h}$$

   - Idea 2 (better): replace limit by local regression intercept,

$$\widehat{\lim_{x \downarrow c} E}[Y|X = x] = \hat{\alpha}$$

$$(\hat{\alpha}, \hat{\beta}) = \underset{a,b}{\operatorname{argmin}} \sum_{i:c<X_i<c+h} \left(Y^1 - a - b \cdot (X_i - c)\right)^2$$

   - More on estimation later in this class!