ECMA 31100: Intro to Empirical Analysis II

RDD

Joe Hardwick

University of Chicago

Winter 2022

Introduction

- In week 1 we discussed estimation of ATE,ATT and ATU under unconfoundedness:

$$y(0), y(1) \perp D|X.$$

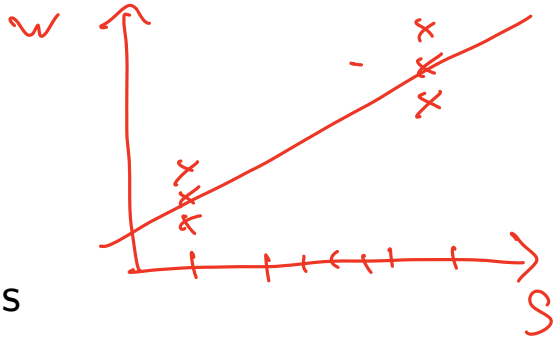
- Saw several representations of the ATE which led to different estimators with different properties.
- For example, also assuming $\mathrm{E}(y(d)|X) = \alpha_d + X'\beta_d$ allowed us to write

$$\mathrm{E}(Y|D,X) = \alpha_0 + \beta_0 X + (\alpha_1 - \alpha_0) D + D \cdot X' (\beta_1 - \beta_0),$$

which gave

$$ATE = (\alpha_1 - \alpha_0) + \mathrm{E}(X)' (\beta_1 - \beta_0).$$

Introduction



- In general identified ATE as

$$ATE = \mathrm{E}(y_1 - y_0) = \mathrm{E}\left[\mathrm{E}(Y|D=1,x) - \mathrm{E}(Y|D=0,x)\right].$$

- Implicitly assumed overlap condition: $0 < \mathrm{P}(D=1|x=x') < 1$ for all x'.
- If for some x', $\mathrm{P}(D=1|x') = 0$, then there are no observations of Y for which $D=1$ and $x=x'$, so $\mathrm{E}(Y|D=1,x=x')$ is undefined.
- Constructed non-parametric estimates of the ATE assuming overlap.

Introduction

- Now consider a case where X is scalar and $D = \mathbf{1}(X \geq c)$.
- Treatment is a deterministic function of the running variable X.
- c is called the cutoff/threshold.
- A regression discontinuity design identifies the effect of the treatment at the cutoff c, and relies on an assumption that the untreated potential outcome mean is continuous at c.

Sharp RDD

- If $D = \mathbf{1}(X \geq c)$, design is called sharp because X determines D.
- Unconfoundedness holds when conditioning on X because of this:

$$y_0, y_1 \perp D|X$$

because D is deterministic after conditioning on X.

- Overlap fails because we never see y_0 and y_1 for the same covariate value x:

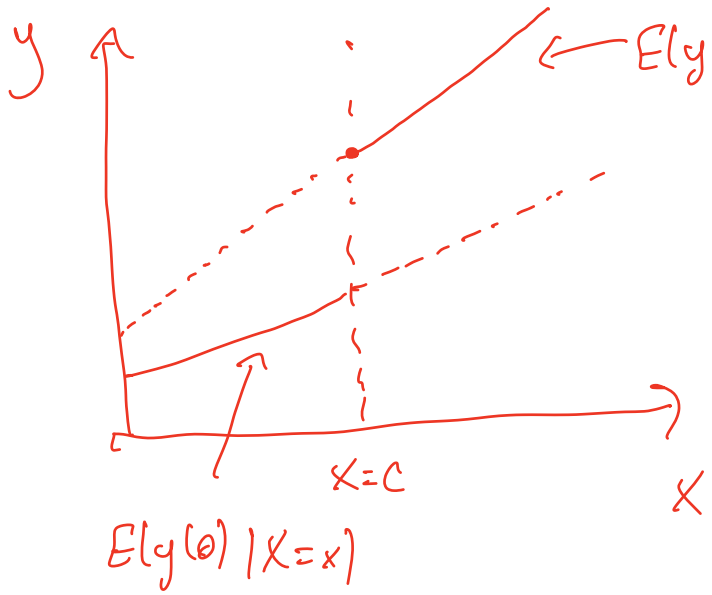
$$\mathrm{P}(D = 1|X = x) = \begin{cases} 1 & x \geq c, \\ 0 & x < c. \end{cases}$$

Sharp RDD

- Assuming linear conditional means allowed us to circumvent the overlap condition in identifying the ATE, provided we could estimate the parameters α_d, β_d.
- This amounts to extrapolation: Estimate α_0, β_0 with i: $X_i \leq c$ and α_1, β_1 with $i : X_i > c$.
- More flexible parametric model could be used.. but doesn't avoid issue of extrapolation.
- With this assumption the ATE is identified because

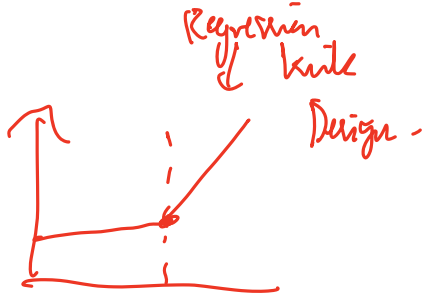
$$ATE = (\alpha_1 - \alpha_0) + \mathrm{E}(X)'(\beta_1 - \beta_0).$$

$ATE = E\left[E(Y|D=1,X) - E(Y|D=0,X)\right]$



Can infer $E(y(1)-y(0)|X=x)$ by extrapolation, and so ATE is identified.

Example: Regression Kink + Discontinuity



- Suppose y is credit limit and x is credit score.
- Suppose only two categories of credit: 'Good' and 'Excellent'.
- Features:
  - Sharp discontinuity in credit limit as soon as individual crosses 'Excellent' boundary (Discontinuity)
  - Each additional unit of credit score improves credit limit more for individuals with excellent credit (Kink)
- Let

$$d_i = \begin{cases} 1 & \text{if individual } x_i \geq 0.5 \\ 0 & \text{if individual } x_i < 0.5 \end{cases}$$

Excellent Credit

⇔ Good Credit

Example: Regression Kink + Discontinuity

- Model:

$$\begin{aligned} Y_i &= \alpha_0 + \beta_0 X_i + D_i (\gamma + \delta X_i) + \epsilon_i \\ &= \alpha_0 + \beta_0 X_i + \gamma D_i + \delta D_i X_i + \epsilon_i. \end{aligned}$$

- When $x < c$: $\mathrm{E}(Y|X = x) = \alpha_0 + \beta_0 x$.
- When $x_i \geq c$: $\mathrm{E}(Y|X = x) = (\alpha_0 + \gamma) + (\beta_1 + \delta) x$.
- Here $\gamma = \alpha_1 - \alpha_0$, $\delta = \beta_1 - \beta_0$.
- δ represents the additional return to credit score for individuals with excellent credit.
- Discontinuity:

$$\mathrm{E}(Y|X = c) - \lim_{x \uparrow c} \mathrm{E}(Y|X = x) = \gamma + \delta \cdot 0.5.$$

Example: Regression Kink + Discontinuity

- Can recenter X to get δ to represent discontinuity:

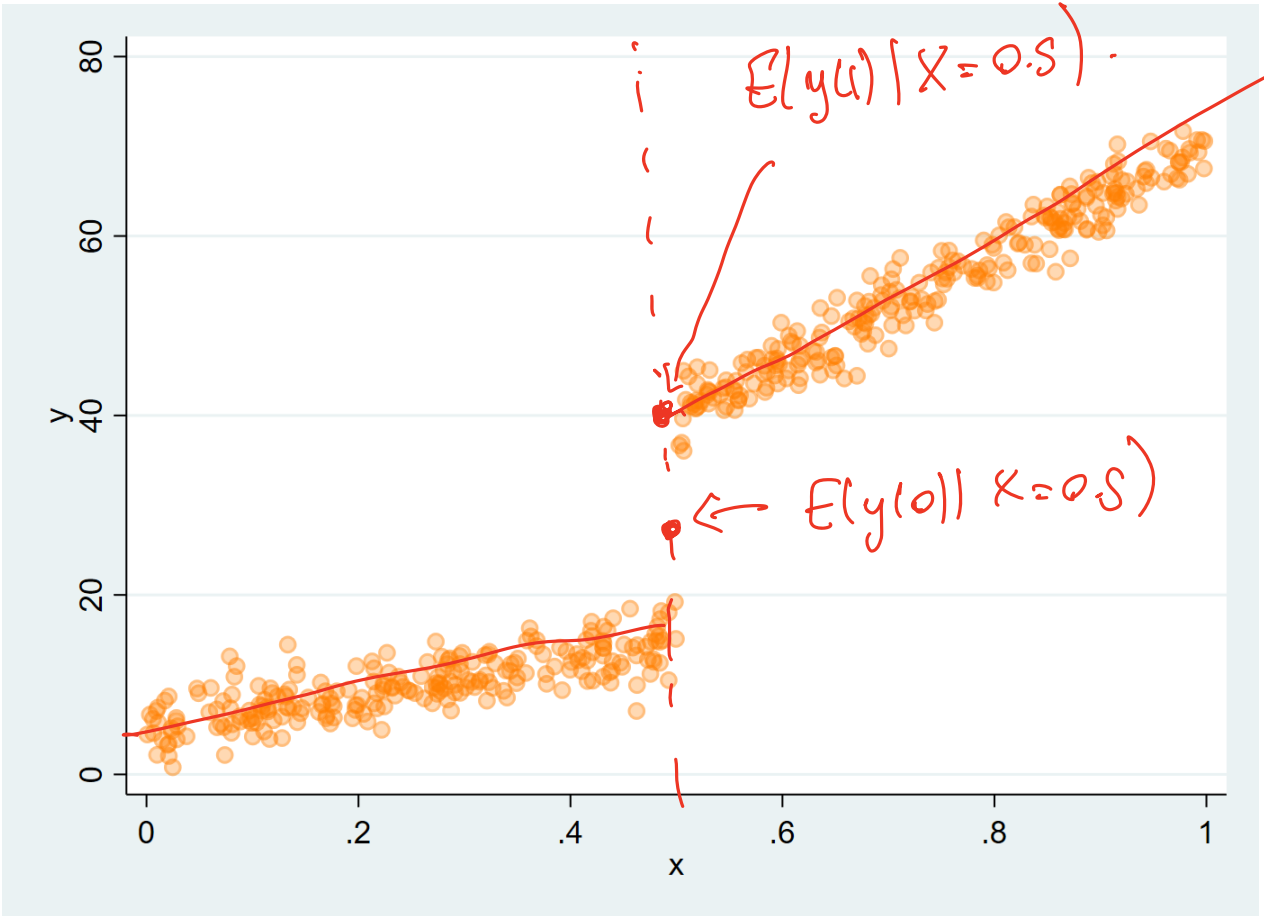
$$Y_i = \alpha_0 + \beta_0 (X_i - c) + \gamma D_i + \delta D_i (X_i - c) + \epsilon_i.$$

- Now

$$\mathrm{E}(Y|X = c) - \lim_{x \uparrow c} \mathrm{E}(Y|X = c) = \gamma.$$

Example

- Credit card limits offered to new customers as a function of credit score:



Sharp RDD

- We can identify the treatment effect at the cutoff with a much weaker assumption:

$$\mathrm{E}\left(y_0|X=x\right) \text{ is continuous at } c.$$

- Then

$$\begin{aligned}\mathrm{E}\left(y_1-y_0|X=c\right) &= \mathrm{E}\left(y_1|X=c\right) - \lim_{x\uparrow c}\mathrm{E}\left(y_0|X=x\right) \\ &= \mathrm{E}\left(Y|X=c\right) - \lim_{x\uparrow c}\mathrm{E}\left(Y|X=x\right).\end{aligned}$$

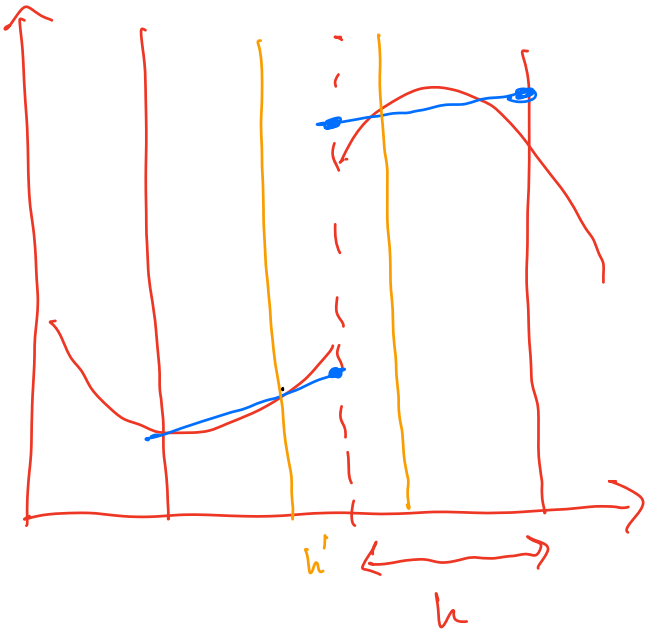
ATE at cutoff

Sharp RDD

- Since we don't want to assume $\mathrm{E}(Y|X=c)$ is continuous at c, we estimate this mean in a neighborhood of c.
- Common approach: Choose a bandwidth h and solve

$$\min_{\alpha_0,\beta_0,\gamma,\delta} \sum_{i=1}^{n} \mathbf{1}(|X_i - c| \leq h)(y_i - \alpha_0 - \beta_0 x_i - \gamma d_i - \delta d_i x_i)^2 .$$

- Called Local Linear Regression. Gives an approximation to conditional mean in a neigborhood of cutoff.



With a smaller bandwidth h', the local linear approximation to the conditional mean will have lower bias but higher variance because there are fewer observations included.

Sharp RDD

- Equivalent to two separate linear regressions on subsamples $\{i : X_i \in [c, c+h]\}, \{i : X_i \in [c-h, c)\}$.
- Split up the OLS minimization problem into two sums:

$$\min_{\alpha_0,\beta_0,\gamma,\delta} \sum_{i=1}^{n} \mathbf{1}\left(|X_i - c| \leq h\right)\left(y_i - \alpha_0 - \beta_0 x_i - \gamma d_i - \delta d_i x_i\right)^2$$

$$\equiv \min_{\alpha_0,\beta_0,\gamma,\delta} \left[\sum_{i:D_i=0}^{n} \mathbf{1}\left(|X_i - c| \leq h\right)\left(y_i - \alpha_0 - \beta_0 x_i\right)^2 \right.$$

$$\left. + \sum_{i:D_i=1}^{n} \mathbf{1}\left(|X_i - c| \leq h\right)\left(y_i - (\alpha_0 + \gamma) - (\beta_0 + \delta) x_i\right)^2 \right]$$

- Can minimize each sum independently since fixing α_0, β_0 leaves γ, δ free.

Bandwidth Choice

- Without linearity of conditional means, h should shrink as $n \to \infty$, otherwise local linear estimator will not necessarily reflect behaviour of $\mathrm{E}(Y|X = x)$ near $x = c$.
- 'Optimal' bandwidth chooses h to minimize an approximation to

$$MSE(h) = \mathrm{E}\left(\hat{\gamma}(h) - \gamma\right)^2.$$

  Turns out that $h = C \cdot n^{-1/5}$ works best.
- This leads to an asymptotic bias term in the asymptotic distribution.

Bandwidth Choice

$h = C n^{-2/5}$.

- Intuitively, larger h leads to higher bias (using observations far from cutoff) but lower variance (larger sample).
- Imbens and Kalyanaraman (2012), Calonico, Cattaneo and Titiunik (2014) propose two different estimates of C.
- CCT account for asymptotic bias term by estimating this bias and incorporating the estimation error in the variance of the resulting estimator of $\mathrm{E}(Y|X = c)$.
- Alternative undersmoothing approach sets $n^{1/5}h \to 0$ which eliminates asymptotic bias.

Failure of Identification

- Manipulation at the cutoff: If some individuals can choose treatment status by ensuring their X value is just above/below cutoff, then individuals to the right and left may not be the same, which might suggest a violation of continuity.
- Several treatments: If there are multiple treatments at the boundary, cannot identify which of them caused the discontinuity unless we essentially assume the second has no effect on average.
- McCrary (2008) proposes a test of a discontinuity in the density of X at c - may suggest manipulation at cutoff. See also Bugni and Canay (2021).

Covariates

Un -

- Confoundedness holds with or without additional covariates W, since D is a function of X.
- Discontinuities in the distribution of pre-determined covariates at the cutoff may also be suggestive of a violation of the continuity assumption, since the change in covariate value might be causing the change in outcome rather than the treatment.
- RDD with outcome replaced by W should show no effect if W responds continuously to X. Null is that $\mathrm{E}(W|X = x)$ is continuous at c.
- Since intuition is about entire distribution of X, Canay and Kamat (2018) propose a more powerful test based on the entire distribution of $W|X$.

$(Y, X) \in \mathbb{R}^2$ want to estimate $E(Y|X=x)$.

When X is discrete, use binning estimator:

$$\hat{m}(x) = \frac{\sum Y_i \, 1(X_i = x)}{\sum 1(X_i = x)}.$$

When X is continuous, we use a neighborhood of c to estimate $m(c)$ using bandwidth h:

$$\hat{m}(c) = \frac{\sum Y_i \, 1[|X_i - c| \le h]}{\sum 1[|X_i - c| \le h]}$$ ← Uniform kernel

$\frac{|X_i - c|}{h} \le 1$

Larger h creates higher bias $\to h \to \infty$ $\hat{m}(c) = \bar{Y}_n$

Smaller h " " variance $\to h \to 0$ $\hat{m}(X_i) \to Y_i$

Low bias but high variance.

In general $\hat{m}(c) = \frac{\frac{1}{nh}\sum Y_i \, k\left(\frac{X_i - c}{h}\right)}{\frac{1}{nh}\sum k\left(\frac{X_i - c}{h}\right)}$ k is called a kernel.

eg. Gaussian kernel: $k(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$.

nh is effective number of observations.

$$\sqrt{nh}\left(\hat{m}(x) - m(x)\right) = \sqrt{nh}\, T_1(x) + \sqrt{nh}\, T_2(x)$$

$\downarrow$ Asymptotic Bias. $\quad$ $\downarrow_d$ $N(0, V)$

nh is effective ~~#~~ of obs.

eg. if $X \sim U[-1,1]$ and use $1\{|X| \le h\}$, a fraction h of observations are included. So eff. sample size is nh.

$$Y_i = m(X_i) + V_i. \quad E(V_i | X_i) = 0 \text{ by construction.}$$

$$Y_i = m(x) + (m(X_i) - m(x)) + V_i.$$

$$\text{Numerator of } \hat{m}(c) = \frac{1}{nh} \sum k\left(\frac{X_i - c}{h}\right) Y_i$$

$$= \frac{1}{nh} \sum k\left(\frac{X_i - c}{h}\right) m(c) + \frac{1}{nh} \sum k\left(\frac{X_i - c}{h}\right)\left(m(X_i) - m(c)\right) + \frac{1}{nh} \sum k\left(\frac{X_i - c}{h}\right) V_i.$$

$$= \hat{f}(c)\, m(c) + \hat{T}_1(c) + \hat{T}_2(c).$$

$\hat{f}$: Kernel density estimate $\frac{1}{nh}\sum K\left(\frac{x_i - c}{h}\right)$

$$\hat{f}(c)\,\hat{m}(c) = \hat{f}(c)\, m(c) + \hat{T}_1(c) + \hat{T}_2(c).$$

$$\hat{m}(c) - m(c) = \frac{1}{\hat{f}(c)}\left(\hat{T}_1(c) + \hat{T}_2(c)\right).$$

It turns out that:

$$\sqrt{nh}\,\hat{T}_2(c) \to^d N\left(0, V(c) f(c)\right)$$

$$\sqrt{nh}\left(\hat{T}_1(c) - \mu h^2 f(c) B(c)\right) \to^d 0. \quad \text{Combining:}$$

$$\sqrt{nh}\left(\hat{m}(c) - m(c) - h^2 \cdot \mu B(c)\right) \to^d N\left(0, \frac{V(c)}{f(c)}\right)$$

since $\hat{f}(c) \to^p f(c)$.

$$AMSE = Var + Bias^2$$

$$= h^4 \mu^2 B(x)^2 + \frac{V}{nh f(x)}$$

Want to set $h^4 \asymp \frac{1}{nh}$ $\quad h^5 \asymp \frac{1}{n}$

$h \asymp n^{-1/5}$.

Sets order of bias^2 = order of variance.

Kernel est. of $m(c)$ essentially solves

$$\hat{m}(c) = \underset{a}{\operatorname{argmin}} \sum k\left(\frac{x_i - c}{h}\right)(y_i - a)^2$$

↑ Local constant estimator.

Local linear estimator (used in ROD) solves:

$$\hat{m}(c) = \underset{a,b}{\operatorname{argmin}} \sum k\left(\frac{x_i - c}{h}\right)(y_i - a - b x_i)^2$$

Fuzzy Design

- Previously discussed case $\mathrm{P}(D = 1|X = x) = \mathbf{1}(x \geq c)$.
- Now assume only that $\mathrm{P}(D = 1|X = x)$ is discontinuous at c.
- Now $Z = \mathbf{1}(X \geq c)$ is an instrument for D.
- Exogeneity satisfied: $y_0, y_1 \perp Z|X$ because Z is a function of X.
- Relevance conditional on $X = x$ requires $Cov(D, Z|X = c) \neq 0$. True since

$$\lim_{x \downarrow c} \mathrm{P}(D = 1|Z = 1, X = x) \neq \mathrm{P}(D = 1|Z = 0, X = x).$$

Fuzzy Design

- Now assume no defiers: $\mathrm{P}\left[D_1 \geq D_0\right] = 1$. Discontinuity at c is therefore a positive jump.
- Assume $\mathrm{E}\left(y_1 - y_0 | X = x, \text{complier}\right)$ is continuous at $x = c$. Same for always and never-takers.
- Assume $\mathrm{P}\left(\text{complier} | X = x\right)$ continuous at $X = c$. Same for always and never-takers.
- Linear IV regression in a neighborhood of c will essentially produce a LATE:

$$\begin{aligned} &\mathrm{E}\left(y_1 - y_0 | X = c, \text{complier}\right) \\ &= \frac{\lim_{x \downarrow c} \mathrm{E}\left(Y | X = x\right) - \lim_{x \uparrow c} \mathrm{E}\left(Y | X = x\right)}{\lim_{x \downarrow c} \mathrm{E}\left(D | X = x\right) - \lim_{x \uparrow c} \mathrm{E}\left(D | X = x\right)}. \end{aligned}$$

Implementation

- Can estimate these conditional means with local linear estimators either side of cutoff.
- Turns out to be equivalent to doing 2SLS with:
- First Stage:

$$D = \pi_0 + \pi_1 Z + \pi_2 (X - c) + \pi_3 Z (X - c) + v$$

- Second Stage:

$$Y = \beta_0 + \beta_1 D + \beta_2 (X - c) + \beta_3 Z (X - c) + u$$

on the subsample $\{i : |X_i - c| \leq h\}$.