Lecture 13: Regression Discontinuity I

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

March 29, 2021

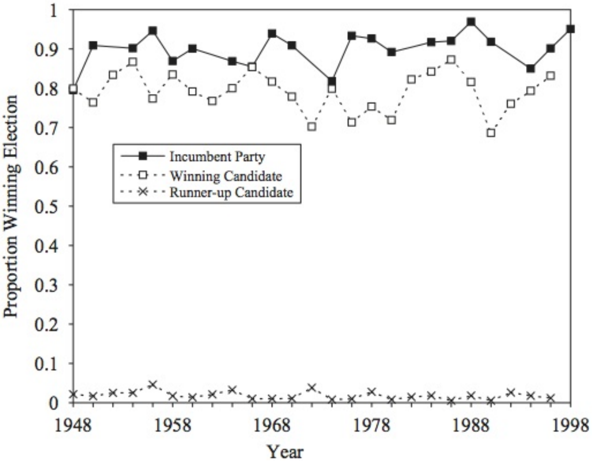
Today:

- Basics of sharp regression discontinuity (RD) designs.

Next session:

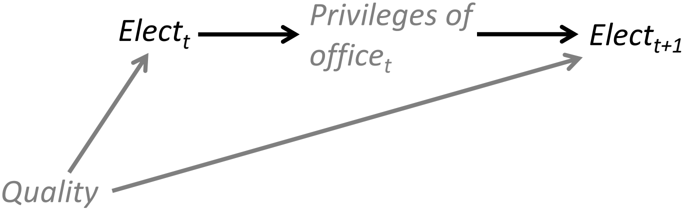
- "Fuzzy" RD.
- "Kink" designs.
- Multiway & geographic RD.
- Threats to validity of RD studies.

Motivating Examples

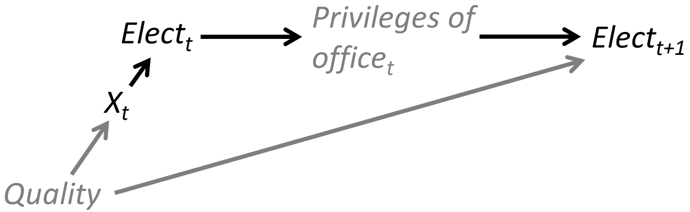


Lee (2008) shows incumbent party & incumbent candidate success rates extremely high. Problem for democratic accountability?

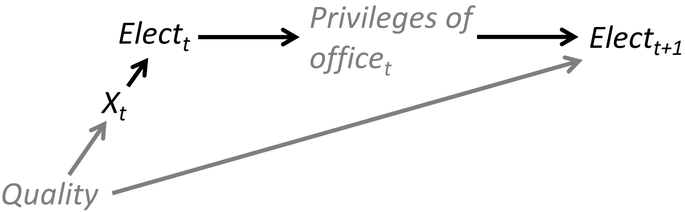
Motivating Examples



Motivating Examples

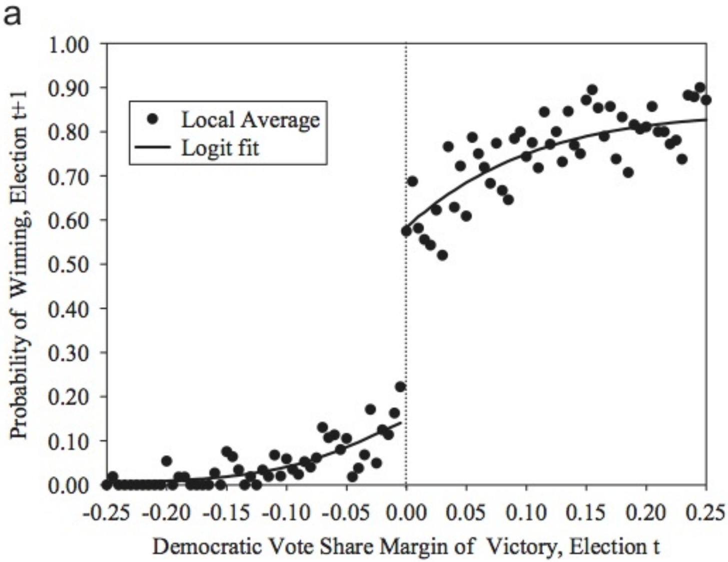


Motivating Examples



Consider X_t = Voteshare$_t$

Motivating Examples



(Lee, 2008)

Motivating Examples

- Some theories of "good governance" propose that incentives in office are important in determining candidate quality and representatives' effort.
- A proposition is that if incentives for office are more financially alluring, then this might improve governance.
- Others argue that barriers to candidacy and limited accountability make such propositions naive and wasteful.
- How can we tell who is right?

Motivating Examples

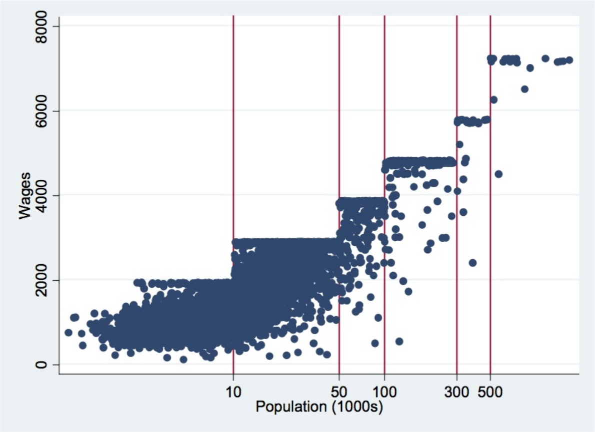
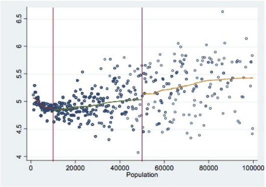


FIGURE 1: LEGISLATORS' SALARIES BY POPULATION

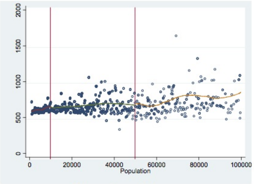
Notes: Figure shows legislators' salaries by population (in log scale). The vertical lines denote the various cutoff points.

(Ferraz and Finan, 2011)

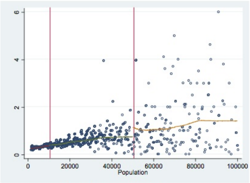
Motivating Examples



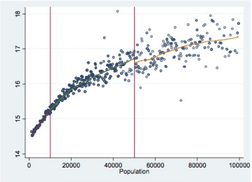
Income per capita (log)



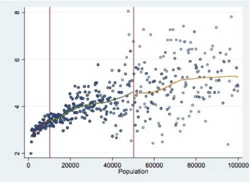
Private Sector Wages



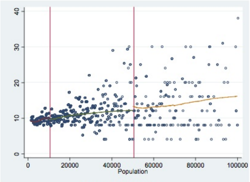
Assistants per legislators



Total Expenditure 2000



Effective Number of Political Parties in 1996 Elections



Hours in session

FIGURE 2: MUNICIPAL CHARACTERISTICS BY POPULATION

Setting

RD identification is based on the idea that in a highly rule-based world, some rules are arbitrary and therefore provide good experiments.
(MHE, p. 251)

Setting

RD identification is based on the idea that in a highly rule-based world, some rules are arbitrary and therefore provide good experiments. (MHE, p. 251)

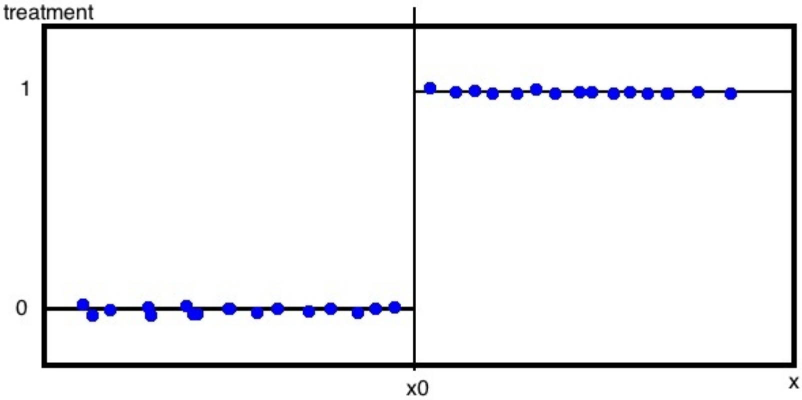
- Suppose a covariate, X_i, that is the basis of a rule for assigning some treatment, $D_i = 0, 1$, such that,

$$D_i = \begin{cases} 1 & \text{if } X_i \geq x_0 \\ 0 & \text{if } X_i < x_0 \end{cases},$$

  where x_0 is some known cut-off point.

- Treatment is then *deterministic* and *discontinuous* in X_i.

Setting



Setting

- Implies *no overlap* in treated and control observations over X_i.
- Antithetical to CIA!
- Anticipates that there will be some modeling involved.
- We nonetheless want to limit dependence on unnecessary assumptions.

Setting

- Start with simple constant effects model:

$$\mathrm{E}\left[Y_{0i}|X_i\right] = f(X_i) \quad \text{and} \quad Y_{1i} = Y_{0i} + \rho,$$

where $f(X_i)$ is a smooth function of X_i (e.g., linear function of X_i, polynomial, sine wave, whatever, so long as it is smooth).

Setting

- Start with simple constant effects model:

$$E\left[Y_{0i}|X_i\right] = f(X_i) \quad \text{and} \quad Y_{1i} = Y_{0i} + \rho,$$

  where $f(X_i)$ is a smooth function of X_i (e.g., linear function of X_i, polynomial, sine wave, whatever, so long as it is smooth).
- We observe,

$$Y_i = f(X_i) + \rho D_i + \eta_i,$$

  where $D_i = 1(X_i \geq x_0)$.

Setting

- Start with simple constant effects model:

$$\mathrm{E}\left[Y_{0i}|X_i\right] = f(X_i) \quad \text{and} \quad Y_{1i} = Y_{0i} + \rho,$$

  where $f(X_i)$ is a smooth function of X_i (e.g., linear function of X_i, polynomial, sine wave, whatever, so long as it is smooth).
- We observe,

$$Y_i = f(X_i) + \rho D_i + \eta_i,$$

  where $D_i = 1(X_i \geq x_0)$.
- ρ is the causal effect.
- D_i is a deterministic function of X_i. No other confounding.
- Causal identification comes from $\mathrm{E}\left[Y_{0i}|X_i\right]$ *smooth* in X_i, but D_i not.
- For identification, *nothing else can change discontinuously* at x_0 other than D_i and outcomes affected by D_i.

Setting

- Now relax constant effects to allow:

$$\mathrm{E}[Y_{0i}|X_i] = f_0(X_i) \quad \text{and} \quad \mathrm{E}[Y_{1i}|X_i] = f_1(X_i),$$

- ($f_0(X_i)$,$f_1(X_i)$ may have different first & higher order derivatives at x_0. Constant effects ruled that out.)

Setting

- Now relax constant effects to allow:

$$\mathrm{E}\left[Y_{0i}|X_i\right] = f_0(X_i) \quad \text{and} \quad \mathrm{E}\left[Y_{1i}|X_i\right] = f_1(X_i),$$

- ($f_0(X_i)$,$f_1(X_i)$ may have different first & higher order derivatives at x_0. Constant effects ruled that out.)
- Weierstrauss approximation theorem: if $f_d(.)$ is continuous, we can approximate it with arbitrary precision with polynomial.

Setting

- Now relax constant effects to allow:

$$\mathrm{E}[Y_{0i}|X_i] = f_0(X_i) \quad \text{and} \quad \mathrm{E}[Y_{1i}|X_i] = f_1(X_i),$$

- $(f_0(X_i), f_1(X_i)$ may have different first & higher order derivatives at x_0. Constant effects ruled that out.)
- Weierstrauss approximation theorem: if $f_d(.)$ is continuous, we can approximate it with arbitrary precision with polynomial.
- Suggests p-th order polynomial approximations,

$$\begin{aligned}
\mathrm{E}[Y_{0i}|X_i] &= \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + \ldots + \beta_{0p}\tilde{X}_i^p \\
\mathrm{E}[Y_{1i}|X_i] &= \alpha + \rho + \beta_{11}\tilde{X}_i + \beta_{12}\tilde{X}_i^2 + \ldots + \beta_{1p}\tilde{X}_i^p,
\end{aligned}$$

where $\tilde{X}_i = X_i - x_0$.

Setting

- Now relax constant effects to allow:

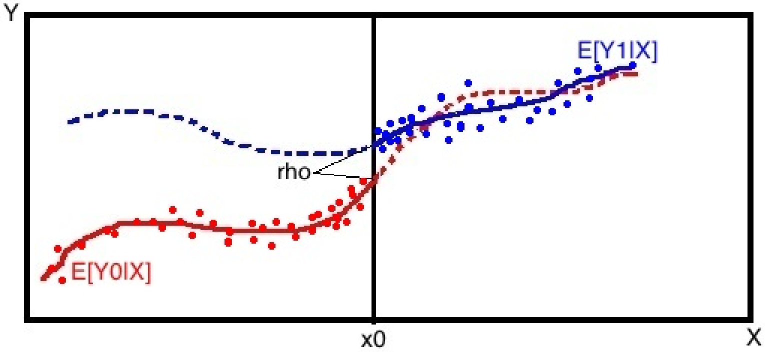
$$\mathrm{E}[Y_{0i}|X_i] = f_0(X_i) \quad \text{and} \quad \mathrm{E}[Y_{1i}|X_i] = f_1(X_i),$$

- ($f_0(X_i)$,$f_1(X_i)$ may have different first & higher order derivatives at x_0. Constant effects ruled that out.)
- Weierstrauss approximation theorem: if $f_d(.)$ is continuous, we can approximate it with arbitrary precision with polynomial.
- Suggests p-th order polynomial approximations,

$$\begin{aligned}\mathrm{E}[Y_{0i}|X_i] &= \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + \ldots + \beta_{0p}\tilde{X}_i^p \\ \mathrm{E}[Y_{1i}|X_i] &= \alpha + \rho + \beta_{11}\tilde{X}_i + \beta_{12}\tilde{X}_i^2 + \ldots + \beta_{1p}\tilde{X}_i^p,\end{aligned}$$

  where $\tilde{X}_i = X_i - x_0$.
- $\rho = \mathrm{E}[Y_{1i}|X_i = x_0] - \mathrm{E}[Y_{0i}|X_i = x_0]$—"treatment effect at x_0."

Setting



- Estimation with interacted regression and centered X_i:

$$
\begin{aligned}
Y_i = & \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + \ldots + \beta_{0p}\tilde{X}_i^p \\
& + \rho D_i + \beta_{11}^* D_i\tilde{X}_i + \beta_{12}^* D_i\tilde{X}_i^2 + \ldots + \beta_{1p}^* D_i\tilde{X}_i^p + \eta_i,
\end{aligned}
$$

where $\beta_{1k}^* = \beta_{1k} - \beta_{0k}$.

Remarks

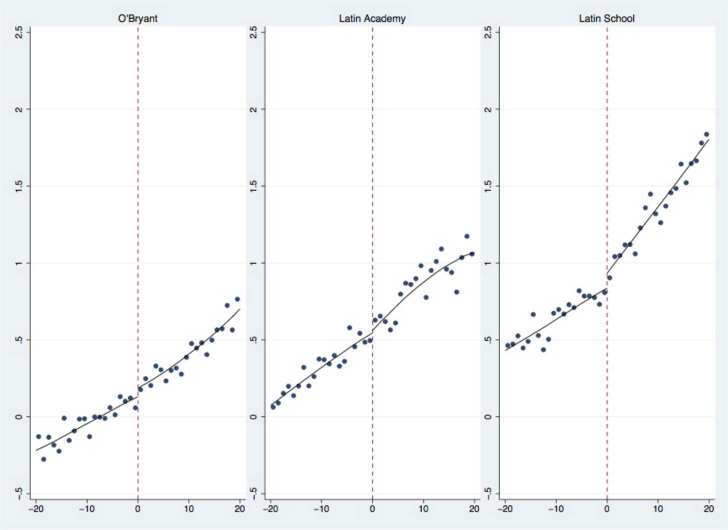


Figure 16. SAT Scores for 7th (2000-2005) and 9th (2001-2006) Grade Applicants in Boston

(Abdulkadiroglu, Angrist, and Pathak, 2011)

Remarks

- Typically no data exactly at x_0.
- $\hat{\rho}$ is a *model based extrapolation*.
- Functional form errors can result in bias.

Remarks

Artifacts of mistaken linearity:

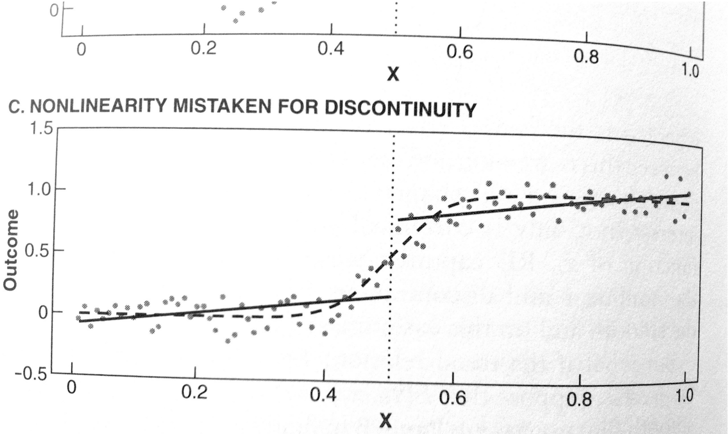


Figure 6.1.1 The sharp regression discontinuity design.

(MHE, p. 254)

Remarks

Artifacts of higher-order polynomials:

Andrew Gelman gave a great example about a year ago on his blog, commenting on a study in PNAS that claimed that China's coal-burning was reducing lifespan by 5 years for half a billion people. Here is the key figure:

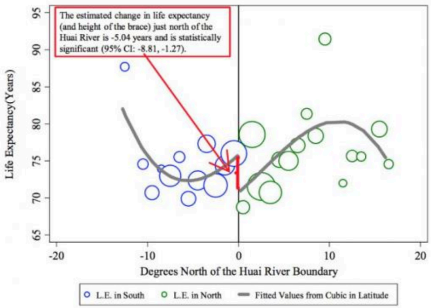


Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

You can see that a cubic is fitted, which results in a statistically significant estimate of -5.5 years. With a linear the estimate is -1.6 years, with a quadratic -1.3 years (neither significant), and with a quartic or quantic, back to -5.4 to -5.6 years and significant.

(From McKenzie, *World Bank Development Impact Blog*, 09/08/2014; cf. Gelman & Imbens, 2017)

Refinements

- Extrapolation errors were due to misspecification.

Refinements

- Extrapolation errors were due to misspecification.
- Consequences of misspecification was exaggerated by letting *data far away from* x_0 determine predictions at x_0.

Refinements

- Extrapolation errors were due to misspecification.
- Consequences of misspecification was exaggerated by letting *data far away from* x_0 determine predictions at x_0.
- We can reduce the potential for such bias by working within a "bandwidth" around x_0, say $[x_0 - \Delta, x_0 + \Delta]$.

Refinements

- Extrapolation errors were due to misspecification.
- Consequences of misspecification was exaggerated by letting *data far away from* x_0 determine predictions at x_0.
- We can reduce the potential for such bias by working within a "bandwidth" around x_0, say $[x_0 - \Delta, x_0 + \Delta]$.
- As bandwidth zeroes in on x_0, we converge on the relevant potential outcomes:

$$\begin{aligned}\lim_{\Delta \to 0} \mathrm{E}\left[Y_i | x_0 < X_i < x_0 + \Delta\right] - \mathrm{E}\left[Y_i | x_0 - \Delta < X_i < x_0\right] \\ = \mathrm{E}\left[Y_{1i} - Y_{0i} | X_i = x_0\right].\end{aligned}$$

- At the limit, we are non-parametrically identified.
- For any fixed Δ we can approximate the CEFs.
- There will be some error. We want to minimize it.

Refinements

- Implementation requires choosing (i) a bandwidth (Δ) and (ii) conditional mean approximations,

$$\hat{\mathrm{E}}\left[Y_i | x_0 < X_i < x_0 + \Delta\right] \quad \text{and} \quad \hat{\mathrm{E}}\left[Y_i | x_0 - \Delta < X_i < x_0\right].$$

- Bias-variance trade-off: less bias as Δ shrinks, but less data too.
- An "optimal" bandwidth would minimize MSE,

$$\mathrm{MSE} = \mathrm{bias}^2 + \mathrm{variance}$$

- Irony: selecting it requires knowing optimal mean approximation, and vice versa!

Refinements

- A first crack at this was Imbens & Kalyanaraman (2012; "IK").
- IK leverage a Porter (2003) result for "edge estimation": *linear* approximation has reliable convergence behavior when bandwidth gets small.
- Thus, IK assume optimal mean approximation will be,

$$\mathrm{E}\left[Y_i | x_0 - \Delta^o < X_i < x_0\right] = \alpha + \beta X_i$$

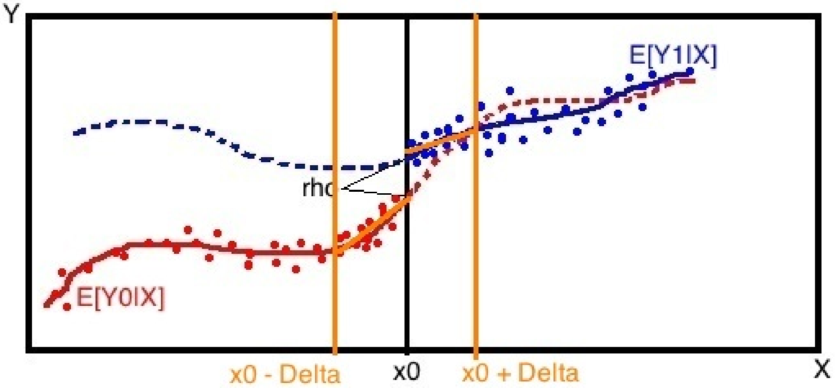
$$\mathrm{E}\left[Y_i | x_0 < X_i < x_0 + \Delta^o\right] = (\alpha + \rho) + (\beta + \gamma) X_i$$

- Estimate with interacted regression and centered X_i:

$$Y_i = \alpha + \rho D_i + \beta \tilde{X}_i + \gamma D_i \tilde{X}_i + \eta_i,$$

  where $\tilde{X}_i = X_i - x_0$ and include only $\{i : X_i \in [x_0 - \Delta^o, x_0 + \Delta^o]\}$.
- Inference for $\hat{\rho}$ follows from usual least squares results.

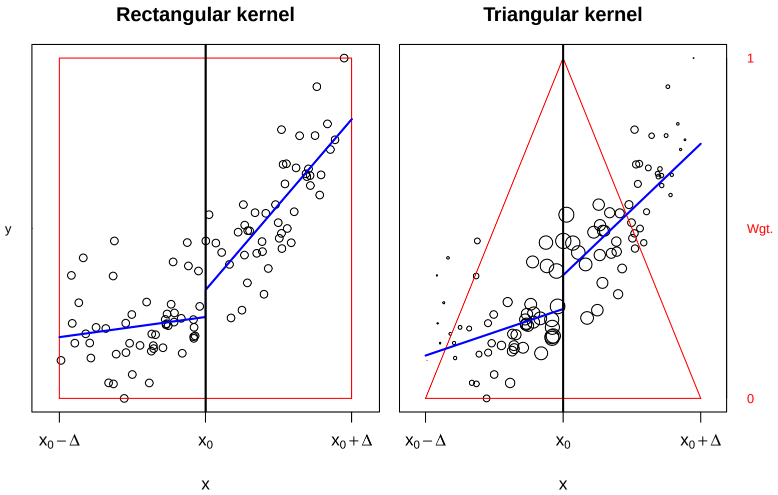
Refinements



Refinements

- Bias can be reduced by down-weighting units far from x_0 (though at a variance cost).
- IK and also Imbens & Lemieux (2009) and Lee & Lemieux (2010) discuss weighting options.
- Results for edge estimation suggest triangular kernel optimal.

Refinements



Refinements

- Start with

$$MSE(\Delta) = \mathrm{E}\left[(\hat{\rho}-\rho)^2\right] = \left(\mathrm{E}\left[\hat{\rho}\right]-\rho\right)^2 + \mathrm{E}\left(\hat{\rho}-\mathrm{E}\left[\hat{\rho}\right]\right)^2$$
$$= \underbrace{\left(\mathrm{E}\left[(\hat{\mu}_+ - \hat{\mu}_-)\right] - (\mu_+ - \mu_-)\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{E}\left((\hat{\mu}_+ - \hat{\mu}_-) - \mathrm{E}\left[(\hat{\mu}_+ - \hat{\mu}_-)\right]\right)^2}_{\text{variance}}$$

where all estimates are within Δ. Want $\Delta^o = \arg\min_\Delta MSE(\Delta)$.

- Key assumptions: iid data, $\tilde{X}_i$ is cts and has mass at outpoint ($f_{\tilde{X}}(0) > 0$), conditional outcome means are three-times differentiable about outpoint, and conditional variance is bounded.
- Define asymptotic MSE (AMSE) in terms of Δ:

$$AMSE(\Delta) = \underbrace{C_1 \Delta^4 \left(\mu_+^{(2)} - \mu_-^{(2)}\right)^2}_{\text{bias}^2} + \underbrace{\frac{C_2}{N\Delta}\left(\frac{\sigma_+^2}{f_{\tilde{X}}(0)} + \frac{\sigma_-^2}{f_{\tilde{X}}(0)}\right)}_{\text{variance}},$$

where C_1 and C_2 are constants that depend on the kernel.

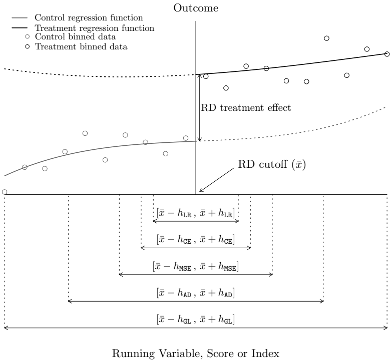
Refinements

- Under this approximation, solve FOC to obtain,

$$\Delta^o = C_K \left(\frac{\sigma_+^2 + \sigma_-^2}{f_{\tilde{X}}(0) \left(\mu_+^{(2)} - \mu_-^{(2)} \right)^2} \right)^{1/5} N^{-1/5}$$

- Implementation:
  - Add regularization term to ensure denom $\neq 0$.
  - $f_{\tilde{X}}(0)$ estimated as share of units near outpoint within pilot bandwidth.
  - σ_+^2 and σ_-^2 estimated as outcome variances within pilot bandwidth.
  - $\mu_+^{(2)}$ and $\mu_-^{(2)}$ estimated off of a polynomial approximation in pilot bandwidth.

Refinements



(Cattaneo & VazquezBare 2016)

- Calonico et al. (2014a, b): refined MSE-approximation—h_{MSE}.
- Calonico et al. (2016): confidence-interval-coverage optimal—h_{CE}.
- Cattaneo et al. (2015): local covariate-balance optimal—h_{LR}.
- Software: `rdrobust` in Stata & R.
- See Cattaneo et al. (2017) for an up-to-date review.

Identification checks

- Recall causal leverage comes from $\mathrm{E}[Y_{0i}|X_i]$ *smooth* in X_i, $D_i = 1(X_i \geq x_0)$ not.
- Identification requires that *nothing else changes discontinuously* at x_0 other than D_i and outcomes affected by D_i.
- (NB: "balanced around x_0"$\Rightarrow$ "smooth around x_0", but "smooth around x_0" $\nRightarrow$ "balanced around x_0")

Identification checks

Imbens and Lemieux (2008) propose a suite of tests:

Identification checks

Imbens and Lemieux (2008) propose a suite of tests:

1. Graphical test of outcome over forcing variable. No visible jump at cut-point or visible jumps away from cut-point undermine credibility of estimated effect.

Identification checks

Imbens and Lemieux (2008) propose a suite of tests:

1. Graphical test of outcome over forcing variable. No visible jump at cut-point or visible jumps away from cut-point undermine credibility of estimated effect.
2. Graphical test and "placebo" RD estimates to test* for jumps in *covariates*. Jumps suggest smoothness violated.

Identification checks

Imbens and Lemieux (2008) propose a suite of tests:

1. Graphical test of outcome over forcing variable. No visible jump at cut-point or visible jumps away from cut-point undermine credibility of estimated effect.
2. Graphical test and "placebo" RD estimates to test* for jumps in *covariates*. Jumps suggest smoothness violated.
3. Graphical and statistical test* for smoothness of density of X_i around x_0. Jumps suggest sorting (McCrary, 2008).

Identification checks

Imbens and Lemieux (2008) propose a suite of tests:

1. Graphical test of outcome over forcing variable. No visible jump at cut-point or visible jumps away from cut-point undermine credibility of estimated effect.
2. Graphical test and "placebo" RD estimates to test* for jumps in *covariates*. Jumps suggest smoothness violated.
3. Graphical and statistical test* for smoothness of density of X_i around x_0. Jumps suggest sorting (McCrary, 2008).
4. Consideration of any jumps in treatment assignment *away* from x_0. These may imply jump at x_0 is being misinterpreted.

*Should use an equivalency test, not test again usual null (Hartman & Hidalgo 2018).

Identification checks

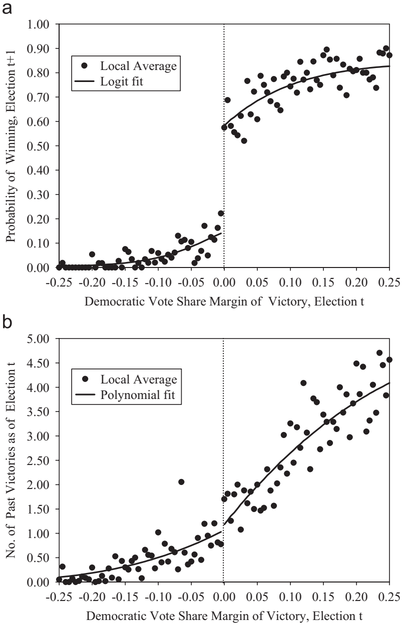
For graphical tests of jumps:

- Using simple binning helps to avoid imposing an allusion of smoothness. A good first cut.
  - Coarsen X_i into bins. Take means within these bins.
- Depending on how you are estimating effects, you can use local linear regression or polynomial regression to refine the tests.

In addition to `rdrobust`, you can use software for local linear regression:

- Stata: `lpoly` function.
  - "`..., kernel(tri) degree(1)...`" is local linear approx with triangular kernel.
- R: `locpol` package and function.
  - `...kernel=TrianK, deg=1,...` is local linear approx with triangular kernel.

Identification checks



(Lee, 2008)

Identification checks

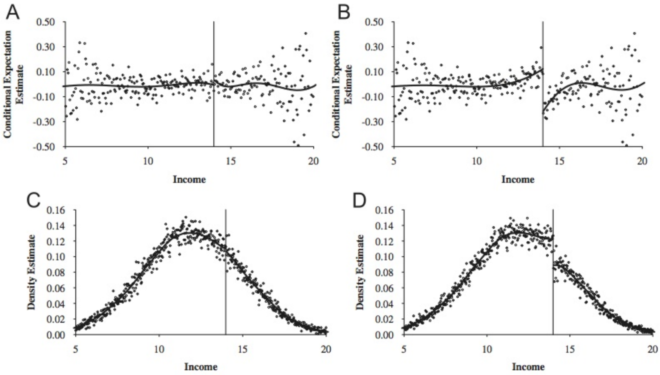


Fig. 2. Hypothetical example: gaming the system with an income-tested job training program: (A) conditional expectation of returns to treatment with no pre-announcement and no manipulation; (B) conditional expectation of returns to treatment with pre-announcement and manipulation; (C) density of income with no pre-announcement and no manipulation; (D) density of income with pre-announcement and manipulation.

(McCrary, 2008)

Remarks

Remarks

- "Treatment effect at x_0" is an odd estimand.
- It is a type of "local average treatment effect" (LATE).

Remarks

- "Treatment effect at x_0" is an odd estimand.
- It is a type of "local average treatment effect" (LATE).
- But typically no units are at exactly x_0. So, it's an effect that applies directly to *no one in particular* (set of measure 0).

Remarks

- "Treatment effect at x_0" is an odd estimand.
- It is a type of "local average treatment effect" (LATE).
- But typically no units are at exactly x_0. So, it's an effect that applies directly to *no one in particular* (set of measure 0).
- Nonetheless smoothness around x_0 means the effect at x_0 is close to what would be the effects of units near x_0.

Remarks

- "Treatment effect at x_0" is an odd estimand.
- It is a type of "local average treatment effect" (LATE).
- But typically no units are at exactly x_0. So, it's an effect that applies directly to *no one in particular* (set of measure 0).
- Nonetheless smoothness around x_0 means the effect at x_0 is close to what would be the effects of units near x_0.
- RD has high "internal validity" but limited external validity (though "better LATE than nothing").
- In an RD study, you should describe covariate values near x_0.
- This will describe the subpopulation for which you are identified.