Regression Discontinuity

Alessandro Caggia

June 2025

Abstract

1. Regression Discontinuity Designs (RDD)

Regression Discontinuity Designs (RDD) are used when the **assignment rule** for treatment is **well known** and based on a **specific continuous variable** called the **running variable**. These designs exploit discontinuities in policy assignment that occur at a known cutoff value of the running variable.

Examples of running variables (assignment variables):

- Income/score threshold to receive financial aid.
- Percentage of votes to win elections.
- Maximum number of students in a class.
- Time (in minutes) applying for a legal permit.

These variables can serve as the dimension along which assignment to treatment changes discontinuously.

Key Assumption: Units on different sides of the threshold are similar.

- The distribution of units along the running variable must be **smooth** around the cutoff.
- That is, units just below and just above the threshold should be comparable.
- Formally: the conditional expectations $E[Y(0) \mid X]$ and $E[Y(1) \mid X]$ must be **continuous** at the cutoff value of X.

Interpretation: For units close to the cutoff, treatment is effectively randomly assigned due to the institutional setup.

- tehy are super similar jsut oen happened ot be on one side and take up tretemnt the other happened to be on the other side
- Around the threshold, the assignment mimics a randomized experiment.
- This allows causal identification under mild continuity conditions.

Two Main Designs:

- **Sharp RDD:** Assignment follows a deterministic rule.
  - Treatment $T = 1$ if and only if running variable X is above the threshold.
  - The probability of treatment jumps from 0 to 1 at the cutoff.
  - The assignment rule is a step function: $\mathbb{P}(T = 1 \mid X)$ is discontinuous and deterministic when plotted against the running variable.
  - Example: financial aid given to students with a score above 80.
- **Fuzzy RDD:** Probability of treatment is discontinuous at the threshold.
  - Assignment does not strictly follow a rule: the probability of treatment changes at the cutoff but not deterministically.
  - The rule is **probabilistic**; not all units above the threshold are treated, and some below may still be treated.
  - This allows for both treated and untreated units on each side of the cutoff.
  - The discontinuity is in *propensity*, not certainty. So the y axis is the probability to get the treatment (proensity score)
  - Possible reasons:
    - 1.Assignment may depend on many **unobserved variables** and you observe just one.
    - 2. You only observe an indicator of eligibility, not compliance. Then risk of **Endogenous assignment:** some units at the threshold may **choose** to comply or not (e.g., eligible but refuse treatment).

1.1. Sharp RDD

Treatment Rule Assume treatment is determined by one covariate, say x_i, according to the rule:

$$D_i = 1[x_i \geq c]$$

This is an indicator function equal to 1 if x_i is greater than or equal to the threshold c, and 0 otherwise. In other words, the rule forces units into treatment or control based on whether they are above or below the cutoff.

Observed Variables We observe:

- The covariate x_i, called the **forcing variable** or **running variable**,
- The treatment assignment D_i,
- The outcome:
$$Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i})$$

This setup defines a standard potential outcomes model where the observed outcome depends on whether the unit is treated ($D_i = 1$) or not.

1.1.1. Potential Outcome Framework

Model for Potential Outcome Means

$$E(Y_0 \mid x) = \mu_0(x), \qquad E(Y_1 \mid x) = \mu_1(x)$$

Conditional Independence Assumption (CIA) Since D is a deterministic function of x, i.e., $D_i = D_i(x_i)$, the **Conditional Independence Assumption** holds:

$$E(Y_g \mid x, D) = E(Y_g \mid x), \quad \text{for } g = 0, 1$$

recall up to now we have sseen it as $(Y_0, Y_1) \perp D \mid X$

- This holds because once we condition on x, which fully determines D, there's no remaining variation in D to explain potential outcomes.

Overlap Assumption Fails Although CIA holds, the overlap assumption does not:

- There is no value of x for which we observe both treated and untreated units.
- If $x < c$, then we only observe control units.
- If $x \geq c$, then we only observe treated units.

- Formally:

$$p(x) = 0 \text{ if } x < c, \qquad p(x) = 1 \text{ if } x \geq c$$

where $p(x)$ is the probability of treatment given x.

Implications

- This lack of overlap makes strategies like regression with controls problematic, as they rely on extrapolation beyond observed data (recall lecture on seleciton on observables).
- Any regression or parametric adjustment would require extrapolating $Y_0(x)$ for treated units or $Y_1(x)$ for control units across the cutoff, which is risky.
- To overcome this, we need strategies to extrapolare y0 (or y1) for units above (below) the cutoff

1.1.2. Homogeneous Treatment Effects

Assume the treatment effect is constant across individuals:

$$Y_{1i} - Y_{0i} = \rho \quad \Rightarrow \quad Y_i = Y_{0i} + \rho D_i$$

Then, taking the expectation conditional on x_i:

$$E(Y_i \mid x_i) = E(Y_{0i} + \rho D_i \mid x_i) = E(Y_{0i} \mid x_i) + \rho E(D_i \mid x_i)$$

$$E(Y_i \mid x_i) = \mu_0(x_i) + \rho \cdot \mathbf{1}[x_i \geq c]$$

note that the lhs is observed in the data: the distribution of yi across each covariate!

Key Identification Assumption:

$$\mu_0(x_i) \text{ is continuous at } c$$

recall that $\mu_0(x_i) := E(Y_{0i} \mid x_i)$.

- this means that limit form above and below of $E(Y \mid X)$ are the same.
- This ensures that any discontinuity in $E(Y_i \mid x_i)$ at $x_i = c$ is entirely due to treatment effect ρ.

Define the limits of the conditional expectation of Y_i approaching the threshold c from the left and from the right:

$$m^-(c) = \lim_{\Delta \to 0} E(Y \mid c - \Delta < x < c) = \lim_{x_i \uparrow c} \mu_0(x_i) + \rho \cdot \mathbf{1}[x_i \geq c] = \mu_0(c)$$

$$m^+(c) = \lim_{\Delta \to 0} E(Y \mid c < x < c + \Delta) = \lim_{x_i \downarrow c} \mu_0(x_i) + \rho \cdot \mathbf{1}[x_i \geq c] = \mu_0(c) + \rho$$

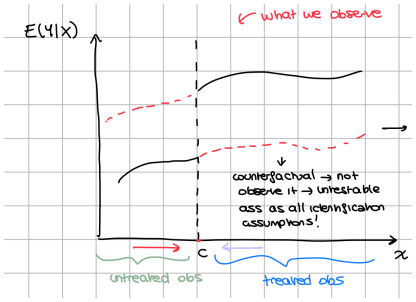
the limits ask: what happens to $E(Y \mid X)$ as i get as close as possible to the cutoff from the one or the other side?

Then the discontinuity in the outcome at the cutoff is:

$$\rho = m^+(c) - m^-(c)$$

- The ATE at a certain value of x is the vertical distance between the two regression curves at that value of x.
- But, we never observe the two regression curves at the same point, only at the cutoff we "almost" observe both curves.
- Thanks to the continuity assumption for the potential outcomes functions, we can recover the ATE at the threshold. MORE in the estiamtion section.

Conclusion: Under the assumption that $\mu_0(x_i)$ is continuous at c, the homogeneous treatment effect ρ is identified as the jump in the conditional expectation of Y at the threshold. The limit of $E(Y \mid X)$ from the left and right of c differ only because of ρ. Counterfactual outcomes $E(Y_0 \mid X \geq c)$ are not observed, making this assumption untestable but necessary for identification.



*see how the distance between the two lines is constant as the treatemtn effect is constant!

Failure of Identification Under Manipulation If individuals can manipulate their running variable x **to obtain treatment** the design is falsified. The problem is that we observe units in the T group that should not be there. we think they belong to such group and have such covariate (eg scored 81 in a test) but truly their score should have been 70. So T and C group are no longer comparable e.g., if those with low Y_0 values push x_i just above $c$then:

- units to the left and the right of the cutoff are no longer similar. Those have an expected untreated coutnerfactua loutcome loer than the trated units. so they drive down the expected value on teh rigth (see image). Than te point is taht for every covariate xi you have adistribution of people with difefrent poetenial outocmes that are later averaged in teh line yo usee. If controls with lower outcomes move to the treated then yo usee also that the outocme mean of the controls goes up!
- $E(Y_0 \mid x) = \mu_0(x_i)$ (the line below!) is no longer continuous at $x = c$.
- The identification assumption is violated.
- We would observe a discontinuity in the mean of Y just above the cutoff not due to ρ, but due to endogenous selection.

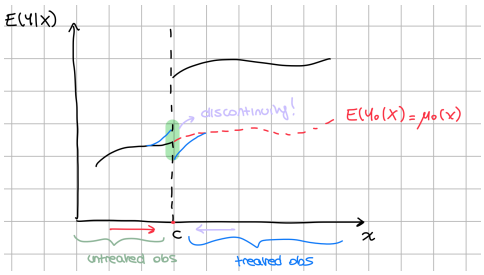


Figure 1. Enter Caption

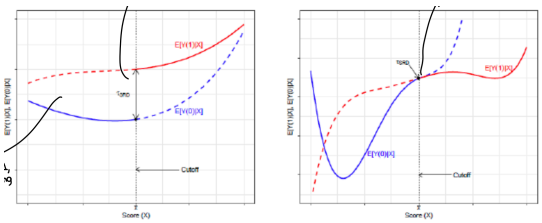
HAVE YOU SEEN ESTIMATORS OF LIMITS? No! So, very difficult in practice to estimate τ!!

1.1.3. interpretation of the Treatment Effect

- Under unrestricted treatment effect heterogeneity, the ATE is not identified.
- **What RDD identifies is a very specific parameter: the treatment effect at $x = c$.**
  - $\rho_c = \mathbb{E}[Y_{1i} - Y_{0i} \mid x_i = c] = \mu_1(c) - \mu_0(c)$
  - This is the ATE for units who are exactly at the cutoff or at the margin of c (marginal idndividuals). SEE THE EQUATIONS ABOVE IN THSI CASE IS NOT ρ is ρ_c of that specific individuals at cutoff c.
  - A very local treatment effect is identified (not LATE, different concept than compliers etc). You could lover the cutoff to get a bit omre T units.

- **Not relevant** to decide whether to switch on or off the policy entirely. Is relevant to udnerstamd whethere scaling up or down the policy would be effective.

SO even if we have heterogeneous treatment effects:



Note that now you need both an assumption on the continuity of $\mu_0(x)$ and $\mu_1(x)$. Why? Before, if $\mu_0(x)$ was continuous, then also $\mu_1(x)$ was (as it was simply a sum with a fixed effect). Now it could be that the treatment effect you are observing is not due to the policy but to the fact that people with a running variable higher than a certain level have far higher counterfactual outcomes. BEFORE RED LINE BY CONSTRUCTION WAS CONTINUS AS BLUE + rho.

1.1.4. Estiamtion

Generic issue

You want units to be at the cutoff, but the closer to the cutoff the less the observations. The choice of the bandwidth leads to a bias-variance trade-off:A) the smaller the sample the more comparable, so the less bias and the more you are closer to the theoretical parameter you want to estimate, but B) low observations lead to high variance and low efficiency due to low observations.

Sharp RDD. Option 1: Global regression using polynomials (not much used anymore)

- reduce variance by usign all the opbservations, limit bias by adding a flexible way to specify the ocndiitonal expectation funciton
- Use the full sample obsevraitons, even those away from the cutoff, and run a regression of Y_i on a constant, D_i, $(x_i - c)$, $D_i(x_i - c)$. If the relaitonship between teh x and the outocme y is easy and trivial and you can easily approximate it with your funcional form, you are fine because you can then spot the discontinuity in y driven by treatment. We use the two strage terms to model the eventual change in splot in the ralaiotnship bwteen the runnin gvariabel and the outcome before and afte tretametn (I mean it coudl happend that the structural relationship between the running avaible and y changes after a certain level of the running variable).

$$Y_i = \alpha + \rho D_i + \beta(x_i - c) + \gamma D_i(x_i - c) + \varepsilon_i$$

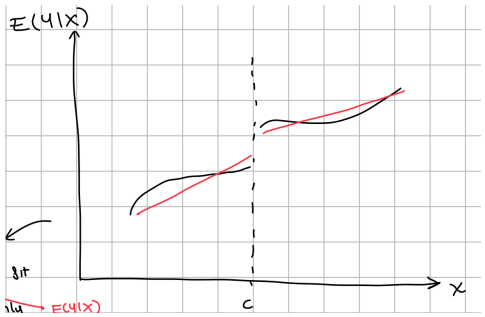
At the cutoff $x_i = c$:

$$\begin{aligned} \text{If } D_i = 0: \quad & Y_i = \alpha \quad \Rightarrow \mathbb{E}(Y_i \mid D_i = 0, x_i = c) = \alpha \\ \text{If } D_i = 1: \quad & Y_i = \alpha + \rho \quad \Rightarrow \mathbb{E}(Y_i \mid D_i = 1, x_i = c) = \alpha + \rho \end{aligned}$$

- We can add a high-order polynomial on $(x_i - c)$ use its squaes, cubes etc.
- The flexibility of a polynomial is important because RDD is sensitive to the functional form of x.

Problems with this method:

- concpwetually bad: it aims at a global optimizztion but we care at mot making errors at cutoff.
- The optimization is global, but we care about errors at the boundaries.
- It can give large weight to points far from the cutoff (if the funcional form is wrong you use observations far away + not well approxinated)
- Estimates are sensitive to the degree of the polynomial.
- if you use thks approach yuo need to show the RD estiamtes are robust to differn orders of the polynomials used to approsimate $E(Y \mid x)$



Sharp RDD. Option 2: Local linear regression

- **Idea**: we care about the ocutcomes for the units close to the cutoff.
- Restrict the sample using a bandwidth h such that $c - h < x_i < c + h$.
- Run the same regression, now locally (Athey and Imbens, 2017).
- Can add polynomials to the local regressions, but not necessary.
- Or run two separate regressions:
  - Y_i on a constant and $(x_i - c)$ for $x_i < c$;
  - Y_i on a constant and $(x_i - c)$ for $x_i \geq c$.
- Estimate for ρ_c is the difference between the intercepts (think that the intercept is the expected value when the $x_i - c = 0$).
- Can add other regressors X_i: regress Y_i on a constant, D_i, $(x_i - c)$, $D_i(x_i - c)$, and X_i.

Problems with this method:

- Small bandwidth → low bias but high variance.
- Large sample size needed near cutoff.
- Must experiment with different h to check robustness.

Two ways of choosing the bandwidth h

1. **Optimal Bandwidth** (Kalyanaraman 2012; Calonico et al. 2014a)
   - Optimal choice depends on:
     - **Second derivative** of regression function at $x = c$ (very large curvatrue means you will make a lot of error through your linear model, so better use smaller bandwith),
     - **Conditional variances** (if outocm evariable has a lot of variance you want more observations to estimate the effects),
     - **Kernel** used in nonparametric regression (the weights you use affect the badwith choice).
2. **Cross-validation** (Imbens and Lemieux, 2008)
   - Run $N - 1$ regressions leaving out one observation and predict its outcome.
   - Choose h that minimizes squared difference between actual and predicted outcomes.
3. **Calonico et al. (2014b, 2017)**: Choose h as the largest window such that pre-treatment differences are not significant between treated and control groups.

Notes on functional form and bias

- In global approach, risk of misspecification and large bias at boundaries.
- In local linear model:

1.2. Fuzzy RDD

Assignment to treatment is not deterministic.

- Imperfect compliance to treatment assignment rule.
- The probability of treatment changes discontinuously at $x = c$, but it does not change from 0 to 1.
- Propensity Score: $P(D = 1 \mid x)$ is the probability of being treaded.
- **Assumption 1:** We assume **continuity** at c for both $\mu_0(x_i)$ and $\mu_1(x_i)$ (allow heterogeneous TE).
- **Assumption 2:** We also need $P(D = 1 \mid x)$ **discontinuous** at c.
- We are going to use this jump in the probability of being treated as an Instrumental Variable for Treatment.

TE repeating the discussion above on the limits

$$\rho(c) = \frac{m^+(c) - m^-(c)}{P^+(D = 1 \mid x = c) - P^-(D = 1 \mid x = c)}$$

This proves identification of $\rho(c)$.
This result has an instrumental variables (IV) flavor:

- Use a variable indicating whether the observation is on the right-hand side (RHS) of the cutoff as an instrument for treatment assignment.
- **First stage:** how much does being on the RHS of the cutoff affect the probability of receiving treatment (denominator).
- **Reduced form:** contrast in mean outcomes between observations just to the right and just to the left of the cutoff.

1.3. Estimation

We can estimate the four terms in the ratio by local linear regression.
We need four regressions:

1. Regress Y_i on a constant and $(x_i - c)$ for $c < x_i < c + h$.
   - The intercept estimates $m^+(c)$.
2. Regress Y_i on a constant and $(x_i - c)$ for $c - h < x_i < c$.
   - The intercept estimates $m^-(c)$.
3. Regress D_i on a constant and $(x_i - c)$ for $c < x_i < c + h$.
   - The intercept estimates $P^+(D = 1 \mid x = c)$.
4. Regress D_i on a constant and $(x_i - c)$ for $c - h < x_i < c$.
   - The intercept estimates $P^-(D = 1 \mid x = c)$.

Here we can choose a bandwidth h once, or choose two separate bandwidths (for numerator and denominator).

IV interpretation

Fuzzy RDD Estimation as IV: The fuzzy RDD estimator from the previous slide is equivalent to an IV estimator. Hahn, Todd, and Van der Klaauw (2001) show that we can estimate ρ_c as the coefficient on D_i in the following IV regression:

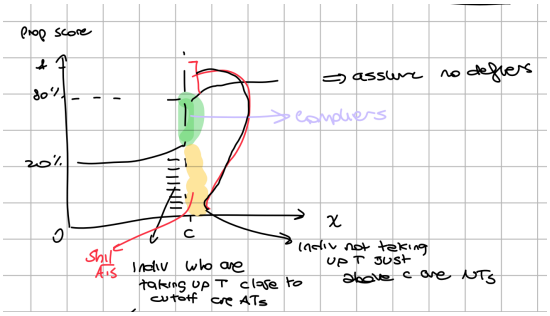
$$Y_i = \alpha + \rho_c D_i + \beta(x_i - c) + \delta D_i(x_i - c) + \varepsilon_i$$

- IVs: use $z_i = \mathbf{1}[x_i \geq c]$ and $z_i \cdot (x_i - c)$ and a Restricted sample with $x_i \in (c - h, c + h)$.
- Use standard IV estimation to recover the LATE at $x = c$. This is the treatment effect for **compliers at** c, i.e., those whose treatment status changes as x_i crosses c.

First Stage Equation:

$$D_i = \alpha + \beta\mathbf{1}(x_i \geq c) + \gamma(x_i - c) + \delta\mathbf{1}(x_i \geq c)(x_i - c) + \varepsilon$$

- $\beta \neq 0$ implies a jump in the probability of treatment at the cutoff (relevance assumption).
- **Randomness/Exogeneity:** $Z \perp Y_0, Y_1$ in this context means $\mathbf{1}(x_i \geq c) \perp Y$ this holds by controlling for x - c in both the first stage and the structural equaiton



* Defiers would lead to a negative shift in the curve similarly as before we identify the ATs when Z = 0, but we focus areound the cutoff.

Estimator Consistency:

- The estimator is consistent for the treatment effect of compliers at $x = c$, **provided** we assume monotonicity (i.e., no defiers).
- monotonicity assuption: If $D_i(k)$ is a function of the cutoff k, we require $D_i(\cdot)$ to be non-decreasing at $k = c$.

Reinterpreting IV Assumptions in Fuzzy RDD:

- **Relevance:** discontinuous jump in the propensity score at the cutoff.
- **Exclusion restriction:** $z_i = \mathbf{1}[x_i \geq c]$ affects Y_i only via D_i. the channel is just that. Note that in the identificaiotn seciton we dfined the ouctome yi as just a fucniton yi(x, D) but not a fucnitn of 1(x>c) (exclusion restriciton is implied).
- **Exogeneity:** z_i behaves like random assignment near c. the asisgnemtn is random. note you could have settings where the asisgnemtn is random but the channel is multiple (recall crime exam...).
- **Monotonicity:** ensures identification of compliers.

Identification:

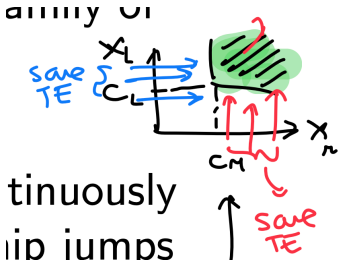
- The denominator of the IV estimator identifies the **proportion of compliers at** c.

RDD as Quasi-Experiment (Lee and Lemieux, 2010):

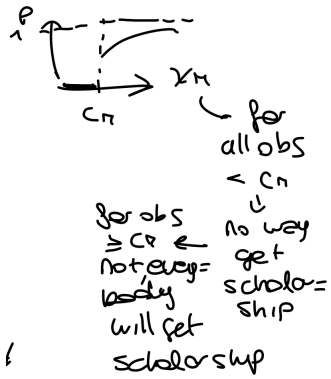
- RDD fails if individuals can precisely manipulate x.
- But, if individuals -even when having some influence- cannot exactly manipulate the assignment variable, the variation in treatment near the threshold is as good as random (bc it still remains contiinous pleas ehtink about it with calm). And we have that evet u idnividual will haev teh same probability of having an x that is just abve or nelow the threshodl (to eb jsut above or below the rhreshold)
- This allows RDD to approximate a randomized experiment. We can test for mean differences in baseline characteristics below and above the threshold.
- Fuzzy RDD is analogous to an experiment with imperfect compliance only ITT is randomized, so we must instrument for treatment.

Extensions

- **Multiple Cutoffs:**
  - example: in one district a party l has 40% of the votes and wins because party k has 38% in another district party l has 20% ad wins because party k has 14%.
  - Pool observations with different cutoffs using a normalized score.
  - Estimation still valid but parameter represents a weighted average of local effects (eg margin of winning election in different districts with more than 2 parties).
- **Multiple Scores:** 54'
  - If treatment depends on exceeding thresholds in multiple tests (e.g. Math and Language), run RD for each score.



  - This yields a family of RD treatment effects.
  - if you observe both the gratdes you ahev a sharp deterministic rule. if you observe just one you haev a one side fuzzy RDD.



- **Regression Kink Design (RKD):**
  - Treatment depends continuously on x, but slope of $D(x)$ changes at cutoff. Equivalent to an RDD on first derivative.
  - RKD estimates treatment effects at the *derivative* level:

$$D = \mathbf{1}\left(x_{\text{math}} \geq c_M \text{ and } x_{\text{lang}} \geq c_L\right)$$

  - Observing only one score implies sharp design (deterministic rule).
  - Suppose the scholarship amount D_i is continuous, but varies with income:
    - Below 30k: you get 5,000 minus 100 per 1,000 income (linear decrease).
    - Above 30k: the decrease becomes steeper, say 300 per 1,000.

    Then:
    - D_i is continuous at 30k.
    - But the slope $\frac{dD}{dx}$ jumps at $x = 30$.

    This is a Regression Kink Design (RKD): *treatment doesn't jump, but its slope changes → kink in the policy rule.* The slope becomes steeper (red line), or even negative: this suggests that additional increases in x are now more associated with increases in treatment probability or possibly decrease it. The slope becomes flatter (blue line), or even negative: this suggests that additional increases in x are now less associated with increases in treatment probability or possibly decrease it.

2. Sumamry

Sharp Regression Discontinuity Design (RDD)

1) Requires SUTVA (Stable Unit Treatment Value Assumption).
2) **Discontinuity assumption:** The probability of treatment jumps at the threshold:

$$P(D = 1 \mid x) \text{ is discontinuous at } c.$$

*discrete and full compliance, discuss at lenght In practice, TEST 1: Check Discontinuity in treatment at threshold

3a) **Continuity assumption:** Units just above and below the cut-off are similar (all below are equiv).

3a1) $\mu_0(x_i)$ is continuous at c. In practice, TEST 2: Check No jump in density of the forcing variable at the threshold (McCrary density test; Null hypothesis: the density of x is continuous at the cutoff (no manipulation). If we fail to reject the null, there is no statistical evidence for manipulation, which supports the validity of the RDD) In practice, TEST 3: Check No Discontinuities in baseline covariates at the threshold

3b) **Conditional Independence Assumption (CIA/ Unconfoundedness):**

$$E(Y_g \mid x, D) = E(Y_g \mid x), \quad \text{for } g = 0, 1$$

* *we are assuming homogeneous TE, CIA garantees continuity at a point and is flat, the other thing guarantees continuity on the full space.

3a and 3b are connected to randomness (local randomization approach).
Test both 3a and 3b No manipulation of the running variable near c (with manipulation both explode)
Under these assumptions, the setup identifies the **ATE** effect, This is not estimating an ITT but directly the ATE you have no intermediate step (take-up), is conceptually different (effect is direct)

Fuzzy Regression Discontinuity Design (RDD)

1) Requires SUTVA.
3) **Continuity assumption:** Both potential outcome functions are continuous at the cutoff (bc we are directly allowing for heterogeneous TE):

$$\mu_0(x_i) \text{ and } \mu_1(x_i) \text{ are continuous at } c.$$

3b **Conditional Independence Assumption** (CIA / Unconfoundedness): Eq

⇒ Under these assumptions, the setup identifies the **Intent-to-Treat (ITT)** effect, as in the IV chapter (actuaoly you do not need 2) for ITT you don't care about the channel being Z = the first stage).

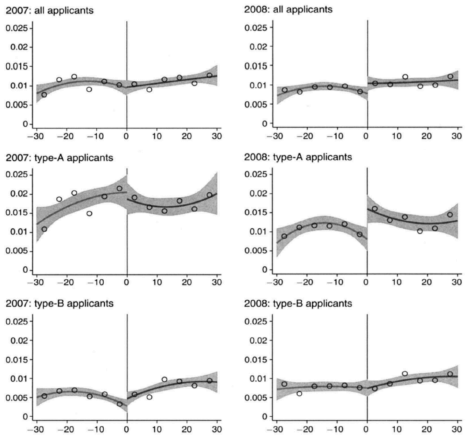
2) **Discontinuity assumption: (p to be treated)**

$$P(D = 1 \mid x) \text{ is discontinuous at } c$$

2a) This is equivalent to the **first stage** (relevance) condition in IV frameworks.

4) **Exclusion restriction:** The running variable Z affects the outcome Y only through the treatment D.
5) **Monotonicity:** No defiers; the direction of treatment assignment is the same for all units at the threshold.

In practice, TEST 4: Placebo check, no discontinuities in the outcome at other placebo cut-offs.



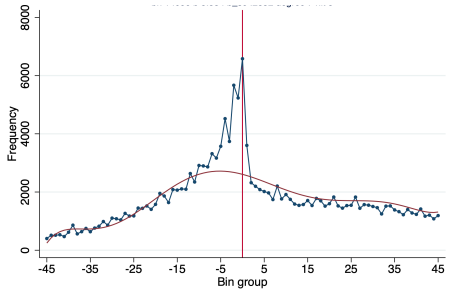
*Common trends are equivalent to the randomness of the instrument.
General point: failure overlap assumption.

Appenidx

About the tests: Great if you have repeated cross section! you can show the discontinuity is there only in the yr of the policy! (Pinotti AER, knocking at heaven's door)

3. Bunching Estimator

The **bunching estimator** identifies behavioral responses (e.g., income elasticity) to non-linear policy rules by detecting excess mass in the density of outcomes near *kink points*, where incentives change discretely (e.g., marginal tax rate jumps). We have a cutoff and we have huge manipulation, we cannot do an RDD, but this is perfect for a bouncing!



Agents choose income z under a piecewise-linear budget constraint. At kink z^* (e.g., where tax rate rises from t_0 to t_1), utility-maximizing individuals may concentrate ('bunch') to avoid higher marginal tax rates. Let:

- $f(z)$: observed income density,
- $h(z)$: counterfactual smooth density (no kink),
- z^*: kink location.

The key identifying assumption is that $h(z)$ is smooth near z^*; deviations of $f(z)$ from $h(z)$ capture behavioral responses.

Estimation Strategy

1. Fit $\hat{h}(z)$ (e.g., polynomial) to $f(z)$ in a bandwidth $[z^* - \Delta, z^* + \Delta]$, omitting a small neighborhood $[z^* - \delta, z^* + \delta]$.
2. Compute excess mass:

$$\hat{B} = \sum_{z \in [z^* - \delta, z^* + \delta]} \left(f(z) - \hat{h}(z) \right)$$

Caveats

- Requires no confounding structural breaks at z^*.
- Sensitive to bandwidth choice, functional form of $h(z)$.