RD Designs with Discrete Running Variables

Northwestern Causal Inference Workshop

Gonzalo Vazquez-Bare

Department of Economics, UC Santa Barbara

August 1, 2024

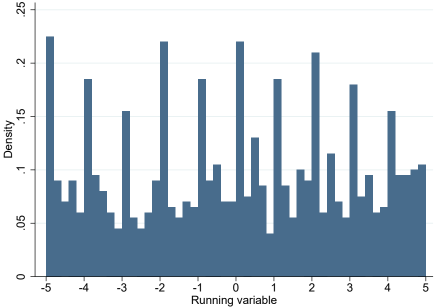
Overview

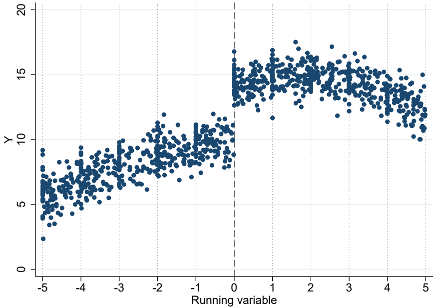
- Discrete random variables take on a finite number of values
  - $X \in \{x_0, x_1, \ldots, x_K\}$ with $\mathbb{P}[X = x_k] > 0$, $\forall\, k$
  - x_k usually called mass points
- RD designs with discrete running variables:
  - Same value of the score shared by multiple units
  - Units may be seen exactly at the cutoff: $\mathbb{P}[X_i = c] > 0$
- Continuity-based methods may not be directly applicable
- But they can still be used in some cases

When can we expect mass points?

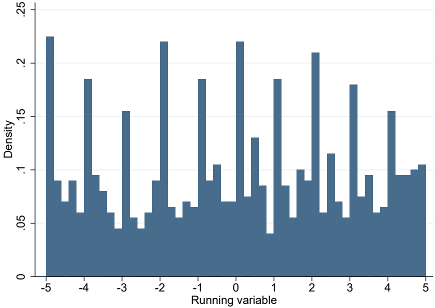
- Case 1: heaping
  - Continuous running variable but with mass points
  - Examples: test scores, birth weight
- Case 2: rounding
  - Underlying continuous variable that is discretized when measured
  - Examples: age in years, income categories
- Case 3: discrete running variable
  - Variable is inherently discrete (e.g. counts)
  - Examples: seats in state Senate, employees in a firm

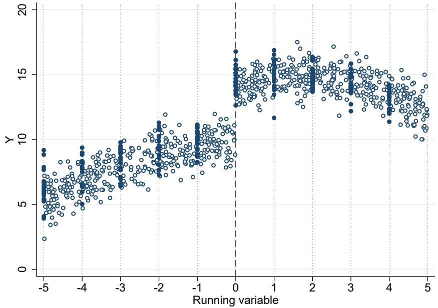
Case 1: heaping



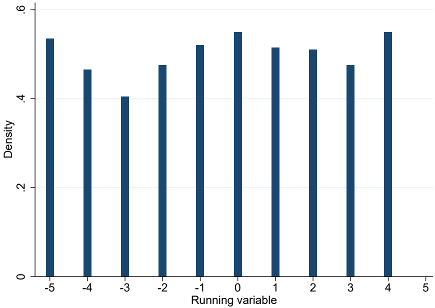


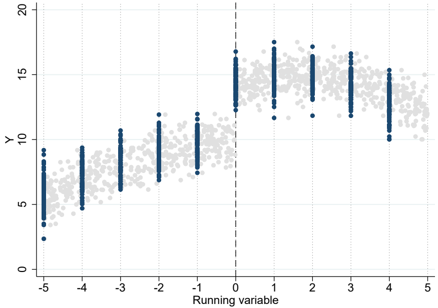
Case 1: heaping



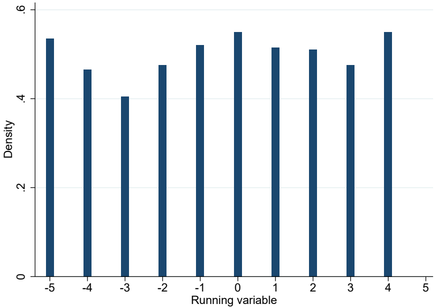


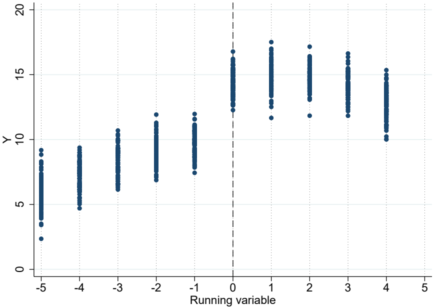
Case 2: rounding





Case 3: discrete RV





Case 1: heaping

- Continuity-based methods may provide a good approximation
  - Many observations close to the cutoff
- Presence of mass points can reduce the effective sample size
  - Local polynomial methods require variation in X to approximate regression functions
  - Adding more observations with the same value of X does not add "useful" variation
- So local polynomials behave as if the total sample size was the number of mass points

Case 1: heaping

- In practice, we can collapse the data at the mass point level
  - Average all observations with the same value of X
  - Avoids this artificial "sample size inflation"
- Can be used as a preferred specification or a robustness check
- See Cattaneo, Idrobo and Titiunik (2024), Section 3

Case 2: rounding

- Underlying latent (unobserved) continuous running variable X^*
- We observe a discrete transformation $X = \phi(X^*) \in \mathcal{S} \subset \mathbb{Z}$
  - E.g. rounding to nearest integer, truncation
- RD effect $\mathbb{E}[Y(1) - Y(0)|X^* = c]$ unidentified (nonparametrically)
  - X^* is unobserved
  - Cannot get arbitrarily close to c using X

Case 2: rounding

- Identification and estimation rely on parametric assumptions
- Even knowing the true model, estimation based on X is inconsistent
  - An example of non-classical measurement error in the regressor
- Dong (2015): bias correction based on parametric assumptions on regression functions and distributional assumptions on X^*

Case 3: discrete RV

- RV (and hence regression functions) inherently discrete
- Continuity-based approach conceptually invalid
  - 3.1 Senate seats? 49.9 employees?
- Comparing units with $X = c$ and $X = c - 1$ can still be a reasonable idea (under further assumptions)
- Can be justified under a local randomization approach (Cattaneo, Frandsen and Titiunik, 2015; Cattaneo, Titiunik and Vazquez-Bare, 2017)

RD with a discrete running variable: summary

- With "many" mass points:
  - Continuity-based methods may provide a reasonable approximation
  - Effective sample size is the number of mass points
  - Consider collapsing the data at the mass point level
- With "few" mass points:
  - Continuity-based methods do not work
  - Local randomization approach may be more appropriate

Digression: RD with time series data

- A policy is introduced at time $t = 0$
- We may want to think of this as an RD:
  - Running variable is time: $t = \ldots, -2, -1, 0, 1, 2, \ldots, T$
  - Treatment indicator $D_t = \mathbb{1}(t \geq 0)$
  - Outcome of interest Y_t
- RD: compare outcomes just before and after policy introduction

Digression: RD with time series data

- This approach has many potential issues
- Running variable is discrete (typically measured in years, months...)
- But local randomization assumptions may be hard to justify
  - Time trends, cycles, seasonality
- All the problems of a before-after estimator
- Diff-in-diff methods may be more credible in this setting

Digression: RD with time series data

- Many issues remain even if t is measured continuously
- Identified effect is the instantaneous effect of the policy
  - Unclear policy relevance
- Asymptotics with $T \to \infty$ do not help
  - Information does not accumulate around the cutoff
  - Large-sample methods (inference, bw selection) not appropriate
- Data most likely not iid (non-stationarity, serial correlation)
- Not a very credible setting for an RD

RD with discrete running variable: empirical example

- Lindo, Sanders and Oreopoulos (2010)
  - Impact of academic probation on future performance in Canada
- Students placed on probation when GPA is below a threshold
  - $X_i =$ GPA
  - $D_i = 1$ if placed on academic probation
  - Slightly different cutoffs across campuses (1.5 and 1.6)
  - Authors normalize running variable in their original analysis