



Regression Discontinuity: Intuition

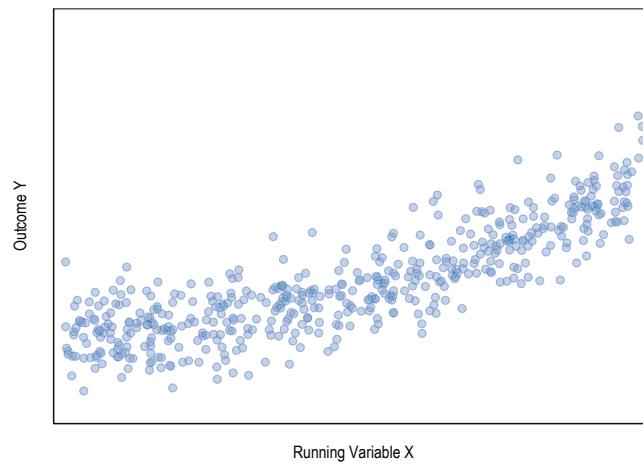
What Is a Regression Discontinuity?

In a **regression discontinuity** (RD) design: treatment status is determined by a precise rule based on a threshold value of a continuous characteristic (the running or forcing variable)

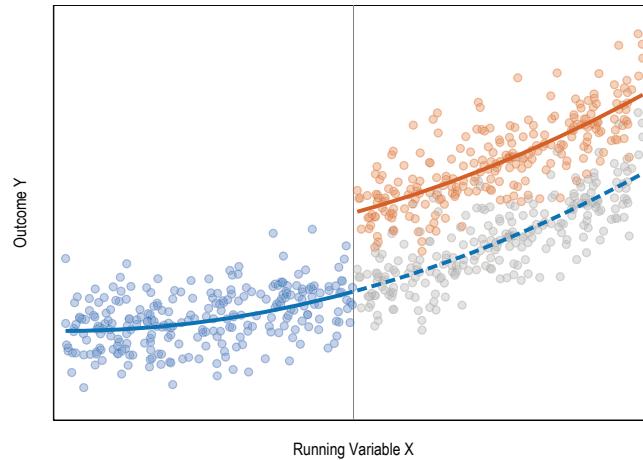
A few examples of potential RDs in development economics:

- **Academic test scores:** admission or scholarships awarded to those with a certain score (e.g. above the threshold for passing), or based on a precise percentile cutoff
- **Household income:** earned-income tax credit for those with incomes below cutoff
- **Poverty indices:** anti-poverty programs are often “proxy-means-tested”
- **Landholdings:** microfinance available to households with less than an acre of land
- **Date of birth:** eligibility for government programs such as preschool/kindergarten, medicare, vaccines, etc. often depends on whether you are above or below an age cutoff
- **Elections:** candidate who gets most (or more than fifty percent of the) votes wins

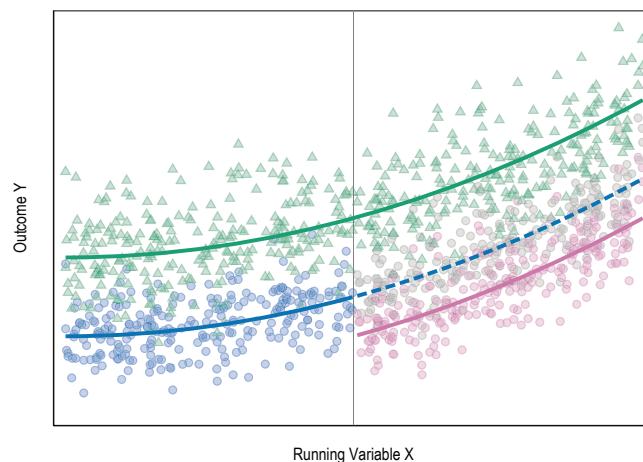
Y_0 Is a Continuous Function of Running Variable X



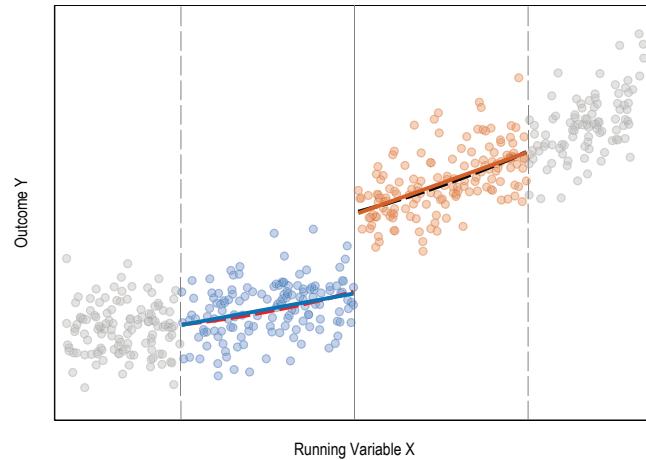
A Jump in $f(X)$ at the Discontinuity Indicates a Treatment Effect



Analogous to Diff-in-Diff, But Without a Comparison Group



With a Narrow Bandwidth, $f(X)$ Should Be Approximately Linear



When Can You Use RD?

RD designs are possible when eligibility is determined by some smooth continuous index, but participants don't have precise control over the index (so they can't select into treatment)

- Individuals with low vs. high values of the eligibility index may look very different, but differences should be small in the neighborhood right around the discontinuity
- Individuals might be able to influence the likelihood of being **near** the discontinuity, but those near the discontinuity can't make sure that they end up on one side vs. the other
 - ▶ Near discontinuity, treatment is "as good as random"

As in a randomized trial, you either have a valid discontinuity or you don't

When Can You Use RD?

For a valid RD, the following must be true:

1. Treatment status determined by a continuous variable (treated \Leftrightarrow if above/below cutoff)
 - ▶ Eligibility index must be smooth, take many values around cutoff
 - ▶ Estimated treatment effect is limit as you approach cutoff
2. Individuals/units/etc. cannot **precisely** manipulate eligibility index
 - ▶ Precise manipulation \Rightarrow selection bias
3. No other programs/treatments can use the same discontinuity

Implementing RD

RD Specification

RD setup:

- Let X_i denote the running variable
- Observations with X_i above cutoff C are treated, those below are not
- Define a re-centered running variable $Z_i = X_i - C$
 - ▶ Let T_i be an indicator for having $Z_i \geq 0$, and hence being treated

The empirical specification for a regression discontinuity design is:

$$Y_i = \alpha + \beta T_i + \theta_{Z_i \geq 0} (Z_i \times T_i) + \theta_{Z_i < 0} (Z_i \times (1 - T_i)) + \varepsilon_i$$

Notice that we use separate slopes above and below the cutoff (could be an interaction)

Bandwidth Selection

Traditionally, most common approach to bandwidth selection is ad hoc: what looks good?

- Choose a range of data small enough such that relationship **looks** linear (in a graph)
- Compare results as you vary the bandwidth, from too small to too large

More recent approaches aim to optimize bias-variance tradeoff

- More data \Rightarrow greater precision, smaller standard errors
- More potential for mis-specification (because continuous functions are locally linear)

Statistical packages (`rd` in Stata) calculate Imbens-Kalyanaraman bandwidth, minimizing MSE

Implementation in Stata

Several different options:

- Run a regression limiting the sample to the appropriate bandwidth
- Use Austin Nichols' `rd` command: `rd Y Z`
 - ▶ Assumes discontinuity occurs at $Z = 0$
- Use Boris Kaiser's `rdcv` command: `rd Y Z, threshold(0) ik`
- Use Sebastian Calonico, Matias Cattaneo, and Rocío Titiunik's `rdrobust`, implementing more robust estimation procedure using different bandwidths for coefficient, standard error

You also need to graph our results (with an RD, it is obligatory)

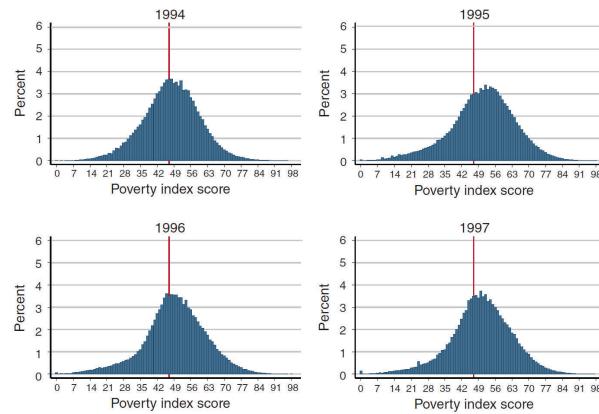
There's Always a Graph

Common elements of RD graphs:

- The running variable is on the x-axis, and the cutoff is indicated
- Observations are binned and bin-level averages are shown in a scatter plot
- The continuous relationship between the running variable and the outcome is shown, always with separate graphs for above vs. below the cutoff for treatment eligibility
 - ▶ Sometimes a linear fit and sometimes a local-polynomial regression
 - ▶ Sometimes with confidence intervals, and sometimes without

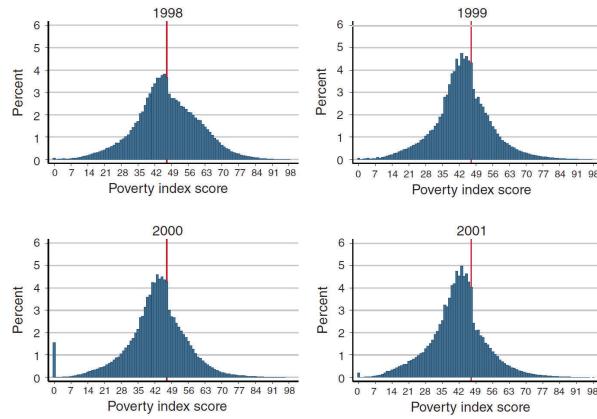
Assessing the Validity of an RD

Q: When Is an RD Not an RD?



Source: Conover and Comacho (2011)

A: After Potential Participants Learn the Allocation Rule



Source: Conover and Comacho (2011)

Defending the Validity of an RD Design

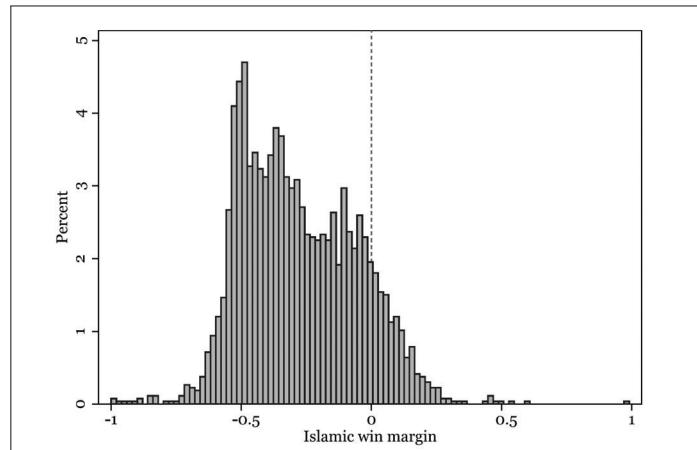
The assumption of no sorting/manipulation can't be tested directly

- Plot the distribution (McCrary (2007) also proposes a test)
- Plot the covariates: pre-treatment characteristics shouldn't jump with treatment

Two approaches empirical approaches:

- Histogram (of either running variable or covariates)
- Local polynomial regression above vs. below (using `lpoly` command in Stata)

Checking for Manipulation: Histogram

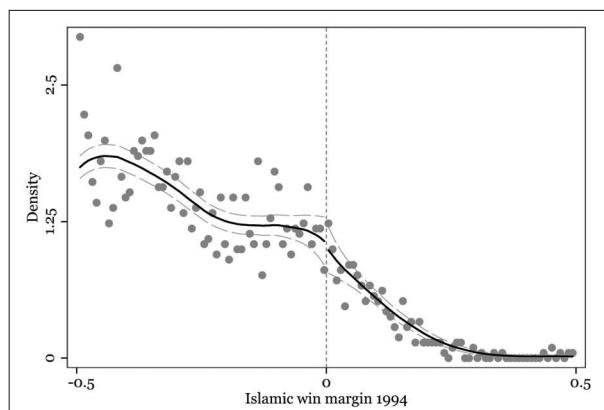


Source: Meyersson (2014)

Economics 523 (Professor Jakielo)

Regression Discontinuity Designs, Slide 27

Checking for Manipulation: Density Plot



Source: Meyersson (2014)

Economics 523 (Professor Jakielo)

Regression Discontinuity Designs, Slide 28

Checking Covariate Balance

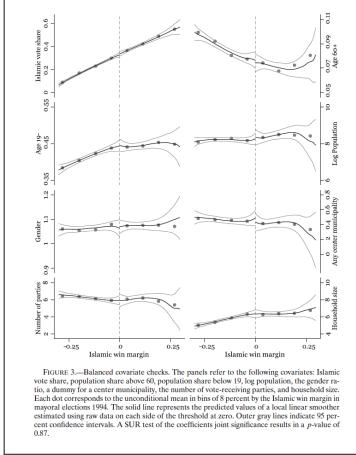


FIGURE 3—Balance diagnostic checks. The panels refer to the following covariates: Islamic vote share, population share above 60, population share below 10, log population, the gender ratio, a dummy for a center municipality, the number of vote-receiving parties, and household size. Each dot corresponds to the unconditional mean in bins of 8 percent by the Islamic win margin in increments of 0.05. The solid line shows the conditional mean estimated by local linear regression, estimated using raw data on each side of the threshold at zero. Outer gray lines indicate 95 percent confidence intervals. A SUR test of the coefficients joint significance results in a p -value of 0.07.

Source: Meyersson (2014)