Lecture 11 – Regression Discontinuity Design

Economics 8379
George Washington University

Instructor: Prof. Ben Williams

Introduction

Estimation in RDD

Identification

RDD implementation

RDD example

Motivation - Differencing

- Regression discontinuity is based on the idea that we sometimes have available an instrument that works “locally” but not globally.

Motivation - Differencing

- Regression discontinuity is based on the idea that we sometimes have available an instrument that works “locally” but not globally.
- Consider the following motivating example...
- The point of this example is to start with something that may be more familiar – a panel data setting.

Motivation - Differencing

- Let D_t be a dummy indicating periods after a policy change, $D_t = \mathbf{1}(t > t_0)$.
- Consider estimating the effect of a policy change using the regression model

$$Y_{it} = \tau D_t + \lambda_t + u_{it}$$

using only entities i that institute the policy change in period t_0

Motivation - Differencing

- We can only do this under restrictions on the λ_t.
  - For example, if $\lambda_{t_0} = \lambda_{t_0+1}$ then $\tau = E(Y_{i,t_0+1}) - E(Y_{i,t_0})$
  - A simple "before-after" estimator doesn't include the λ_ts at all so that $\tau = E(Y_{it} \mid D_t = 1) - E(Y_{it} \mid D_t = 0)$.

Motivation - IV

- Let $Z_t = t$.
- We can use this as an instrument and define various Wald estimators:

$$\frac{E(Y_{it} \mid Z_t = z') - E(Y_{it} \mid Z_t = z)}{E(D_t \mid Z_t = z') - E(D_t \mid Z_t = z)}$$

- If time is exogenous in the model then $Y_{it} = \lambda + \tau D_t + u_{it}$ and all of the Wald estimators with a nonzero denominator are consistent estimators for τ.
- 2SLS will be consistent too.

Motivation - IV

- If $Y_{it} = \lambda_t + \tau D_t + u_{it}$ then time is not exogenous.
- But

$$\frac{E(Y_{it} \mid Z_t = t_0 + 1) - E(Y_{it} \mid Z_t = t_0)}{E(D_t \mid Z_t = t_0 + 1) - E(D_t \mid Z_t = t_0)} = \tau + \lambda_{t_0+1} - \lambda_{t_0}$$

Regression discontinuity.

- Consider estimation of the causal effect of D_i on Y_i. (no time subscripts now)
  - Suppose $Pr(D_i = 1 \mid Z_i = z) = 1$ for $z > z_0$ and $Pr(D_i = 1 \mid Z_i = z) = 0$ for $z \leq z_0$
  - The *sharp* RD estimand is

$$\tau := \lim_{z \to z_0^+} E(Y \mid Z = z) - \lim_{z \to z_0^-} E(Y \mid Z = z)$$

Regression discontinuity

- Consider estimation of the causal effect of D_i on Y_i. (no time subscripts now)
  - Suppose $Pr(D_i = 1 \mid Z_i = z)$ is discontinuous as a function of z at z_0.
  - The *fuzzy* RD estimand is

$$\tau := \frac{\lim_{z \to z_0^+} E(Y \mid Z = z) - \lim_{z \to z_0^-} E(Y \mid Z = z)}{\lim_{z \to z_0^+} Pr(D = 1 \mid Z = z) - \lim_{z \to z_0^-} Pr(D = 0 \mid Z = z)}$$

Regression discontinuity

- When this potentially works
  - There is a substantial “jump” in treatment probabilities $Pr(D_i = 1 \mid Z_i = z)$.
  - Z_i is exogenous near z_0 – allows for endogeneity but not bunching
  - Sufficient data near z_0.
  - Y would vary continuously with Z at z_0 if there is no causal effect.

Examples

TABLE 5 (*continued*)
REGRESSION DISCONTINUITY APPLICATIONS IN ECONOMICS

Study	Context	Outcome(s)	Treatment(s)	Assignment variable(s)
Jacob and Lefgren (2004b)	Elementary schools, Chicago	Test scores	Summer school attendance, grade retention	Standardized test scores
Leuven, Lindahl, Oosterbeek, and Webbink (2007)	Primary schools, Netherlands	Test scores	Extra funding	Percent disadvantaged minority pupils
Matsudaira (2008)	Elementary schools, Northeastern United States	Test scores	Summer school, grade promotion	Test scores
Urquiola (2006)	Elementary schools, Bolivia	Test scores	Class size	Student enrollment
Urquiola and Verhoogen (2009)	Class size sorting- RD violations, Chile	Test scores	Class size	Student enrollment
Van der Klaauw (2002, 1997)	College enrollment, East Coast College	Enrollment	Financial Aid offer	SAT scores, GPA
Van der Klaauw (2008a)	Elementary/middle schools, New York City	Test scores, student attendance	Title I federal funding	Poverty rates
Labor Market				
Battistin and Rettore (2002)	Job training, Italy	Employment rates	Training program (computer skills)	Attitudinal test score
Behaghel, Crepon, and Sedillot (2008)	Labor laws, France	Hiring among age groups	Tax exemption for hiring firm	Age of worker
Black, Smith, Berger, and Noel (2003); Black, Galdo, and Smith (2007b)	UI claimants, Kentucky	Earnings, benefit receipt/duration	Mandatory reemployment services (job search assistance)	Profiling score (expected benefit duration)

Examples

- Example 1. Thistlethwaite and Campbell (1960)
    - Scholarships awarded based on a test score cutoff.
- Example 2. Lee (2008)
    - Incumbency of a particular political party is determined by winning a plurality of votes in the previous election.

Extensions

- We will focus on this version of RDD but we will also briefly consider:
  - if D is not binary
  - if there are multiple thresholds (multiple z_0's)

Estimation

- Consider first estimation of

$$\lim_{z \to z_0^+} E(Y \mid Z = z) - \lim_{z \to z_0^-} E(Y \mid Z = z)$$

  - This is the sharp RDD estimator, but also an intent-to-treat estimator in the fuzzy RDD.
  - In the fuzzy RDD, this can be combined with a similar estimator for the first stage.

Estimation

- Standard estimators (parametric and nonparametric) of $E(Y \mid Z = z)$ *assume* that it is continuous.
- So, in a certain sense, we have to presuppose a jump in order to look for one.
- We want methods that are robust enough that a jump will only be estimated if there is indeed a discontinuity in $E(Y \mid Z = z)$.

Estimation

- A general estimation framework:
- Separate regression equations on the left and right of the threshold:

$$Y_i = \alpha_l + f_l(Z_i - z_0) + \varepsilon_i, \quad z \leq z_0$$
$$Y_i = \alpha_r + f_r(Z_i - z_0) + \varepsilon_i, \quad z > z_0$$

where $f_l(0) = f_r(0)$. Then $\tau = \alpha_r - \alpha_l$.

- Typically these are pooled:

$$Y_i = \alpha_l + (\alpha_r - \alpha_l) D_i + f_l(Z_i - z_0) + D_i \left(f_r(Z_i - z_0) - f_l(Z_i - z_0)\right) + \varepsilon_i$$

where $D_i = \mathbf{1}(Z_i \geq z_0)$.

Linear model

- If f_l and f_r are linear, we have

$$Y_i = \beta_0 + \beta_1(Z_i - z_0) + \tau D_i + \gamma(Z_i - z_0)D_i + \varepsilon_i$$

- Imposing a constant slope:
  - the regression equation is then

$$Y_i = \beta_0 + \beta_1(Z_i - z_0) + \tau D_i + \varepsilon_i$$

  - this is only justified by the fact that it improves efficiency *if* the slope is really constant

Polynomial model

- Polynomial of degree q can be implemented by
  - Define

$$f_r(u) = \beta_{r1} u + \ldots + \beta_{rq} u^q$$
$$f_l(u) = \beta_{l1} u + \ldots + \beta_{lq} u^q$$

  - Then construct the pooled regression equation,

$$Y_i = \alpha_l + (\alpha_r - \alpha_l) D_i + f_l(Z_i - z_0) + D_i \left(f_r(Z_i - z_0) - f_l(Z_i - z_0) \right) + \varepsilon_i$$

Estimation

- Problems with “parametric” models, and solutions.
  - Wrong functional form can bias estimates severely:
    - the polynomial order can be chosen via cross-validation, making this a nonparametric estimate
    - note however, that series estimators can be misbehaved near boundaries because it is a global estimator
  - Identification is local so estimation should be local:
    - a practical solution has been to estimate the above regressions in a window around the cutoff

Estimation

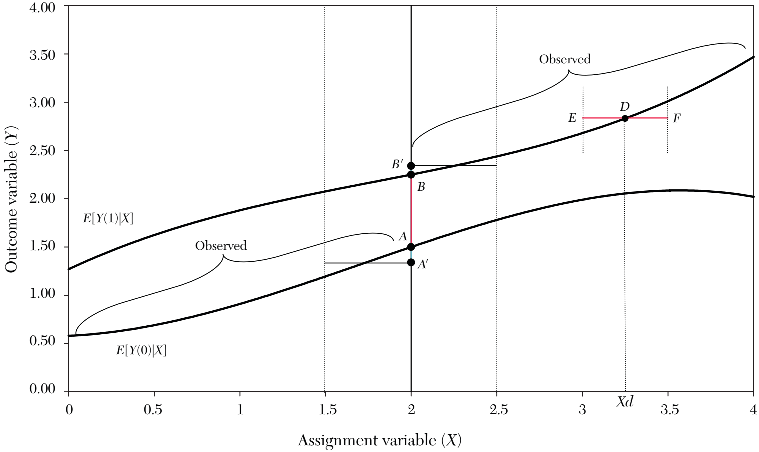


Figure 2. Nonlinear RD

Nonparametric regression

- Nonparametric regression
  - Estimate two nonparametric regression estimators, $\hat{\mu}_1(z)$ and $\hat{\mu}_0(z)$, using subsets of data with $Z_i \geq z_0$ and $Z_i < z_0$, respectively.
  - Then $\hat{\tau} = \hat{\mu}_1(z_0) - \hat{\mu}_0(z_0)$
  - An important problem again arises due to boundary issues – some nonparametric estimators work well for z in the interior of the support but not for z on the boundary of the support.
  - The choice of bandwidth is also a tricky issue: Calonico, Cattaneo, and Titiunik (2014)

Nonparametric regression

- Kernel regression (Nadaraya-Watson) is a local weighted average:

$$\hat{\mu}_1(z_0) := \frac{\sum_{i:Z_i \geq z_0} \frac{1}{h} K\left(\frac{Z_i - z_0}{h}\right) Y_i}{\sum_{i:Z_i \geq z_0} \frac{1}{h} K\left(\frac{Z_i - z_0}{h}\right)}$$

  and analogously for $\hat{\mu}_0(z_0)$
- The window size, h, is called the *bandwidth*.

Nonparametric regression

- Local linear regression minimizes a weighted sum of squares

$$\sum_{i:Z_i \geq z_0} \frac{1}{h} K\left(\frac{Z_i - z_0}{h}\right)(Y_i - a_0 - a_1(Z_i - z_0))^2$$

and analogously for $\hat{\mu}_0(z_0)$

- If $K(u) = \mathbf{1}(|u| \leq 1)$ (a rectangular kernel) then these simply amount to estimating an average or a linear regression in a small window to the right and left of z_0.

Estimation

- Local linear regression is preferred to kernel regression as the latter is more biased on the boundary.

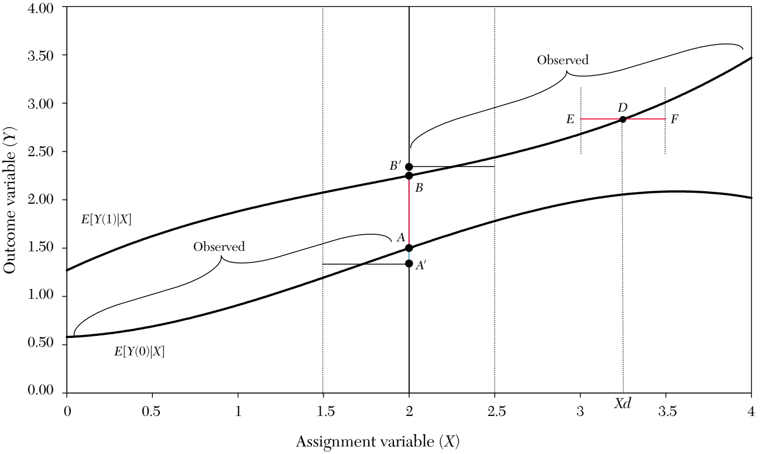


Figure 2. Nonlinear RD

Nonparametric regression

- Nonparametric regression estimators such as these are always biased.
- Bias converges to 0 as $h \to 0$.
- But variance grows to ∞ as $h \to 0$.
- The bandwidth is chosen to balance bias and variance – a slightly larger h is needed to get valid standard errors (over-smoothing).
- Calonico, Cattaneo, and Titiunik (2014) have developed a method for choosing h optimally for RDD.

Fuzzy design

- The fuzzy RD estimand is a Wald estimator:

$$\frac{\lim_{z\to z_0^+} E(Y \mid Z = z) - \lim_{z\to z_0^-} E(Y \mid Z = z)}{\lim_{z\to z_0^+} Pr(D = 1 \mid Z = z) - \lim_{z\to z_0^-} Pr(D = 0 \mid Z = z)}$$

- This suggests estimating the triangular system

$$\begin{aligned} Y_i &= \alpha_l + (\alpha_r - \alpha_l) D_i + f_l(Z_i - z_0) \\ &\quad + T_i \left(f_r(Z_i - z_0) - f_l(Z_i - z_0) \right) + \varepsilon_i \\ D_i &= \gamma_l + (\gamma_r - \gamma_l) T_i + g_l(Z_i - z_0) \\ &\quad + T_i \left(g_r(Z_i - z_0) - g_l(Z_i - z_0) \right) + \nu_i \end{aligned}$$

where D_i denotes treatment and $T_i = \mathbf{1}(Z_i \geq z_0)$.

Fuzzy design

- 2SLS is equivalent to the ratio of reduced form to first stage coefficient.
- May be more efficient to use different polynomial degree and different bandwidths for two stages.
- Generally advised to use same degree and bandwidths however.
  - simpler
  - 2SLS std errors are valid

Identification

- Hahn, Todd, and van der Klaauw (2001):
  1. Suppose $Y_i = \alpha_i + \beta D_i$, which implies constant treatment effects, $Y_{1i} - Y_{0i} = \beta$.
     - In addition, assume that $E(\alpha_i \mid Z_i = z)$ is continuous in z at z_0.
     - Then $\tau = \beta$ (RDD identifies the constant treatment effect)
     - This is for sharp *and* fuzzy design.

Identification

- Hahn, Todd, and van der Klaauw (2001):
  2. Suppose $Y_i = \alpha_i + \beta_i D_i$ and consider the sharp design.
     - In addition, assume that $E(\alpha_i \mid Z_i = z)$ and $E(\beta_i \mid Z_i = z_0)$ are both continuous in z at z_0.
     - Then $\tau = E(\beta_i \mid Z_i = z_0)$
     - Lee and Lemieux (2010) point out that if $\beta_i = \beta(U_i)$ then

$$E(\beta_i \mid Z_i = z_0) = \int \beta(u) \frac{f_{Z_i|U_i}(z_0 \mid u) f_{U_i}(u)}{f_{Z_i}(z_0)} du$$

     - This is a weighted average of treatment effects – weight for u is zero if z_0 is not in the support of $Z_i \mid U_i = u$.
     - Consider for example, the effect of retirement on health using an age cutoff.

Identification

- Hahn, Todd, and van der Klaauw (2001):
  2. Suppose $Y_i = \alpha_i + \beta_i D_i$ and consider the fuzzy design.
     - In addition, assume that $E(\alpha_i \mid Z_i = z)$ and $E(\beta_i \mid Z_i = z_0)$ are both continuous in z at z_0.
     - Also, assume that D_i is independent of β_i conditional on Z_i near z_0.
     - Then $\tau = E(\beta_i \mid Z_i = z_0)$

Identification

- The assumption that D_i is independent of β_i conditional on Z_i near z_0 is a strong assumption.
- In the Roy model, is $E(U_{1i} - U_{0i} \mid D_i, Z_i = z) = E(U_{1i} - U_{0i} \mid Z_i = z)$ for z near z_0?
  - No, and it is not “easier to satisfy” because it is only assumed locally.
- The assume is trivially true in the *sharp* design though!

Identification

- Recall that if
  - "monotonicity": $D_i(z_2) \geq D_i(z_1)$ for all i (or $D_i(z_1) \geq D_i(z_2)$ for all i)
  - $(Y_{1i}, Y_{0i}, \{D_i(z)\})$ independent of Z_i

  then IV estimates the LATE parameter:

  definition. $LATE = E(Y_1 - Y_0 \mid D(z_2) = 1, D(z_1) = 0)$

Identification

- Hahn, Todd, and van der Klaauw (2001):
  3. Suppose $Y_i = \alpha_i + \beta_i D_i$ and consider the fuzzy design.
     - In addition, assume that $E(\alpha_i \mid Z_i = z)$ is continuous in z at z_0.
     - Also, assume that $D_i(z)$ and β_i are jointly independent of Z_i near z_0.
     - and that $D_i(z_0 + e) \geq D_i(z_0 - e)$ for all sufficiently small e
     - Then $\tau = \lim_{e \to 0} E(\beta_i \mid D_i(z_0 + e) = 1, D_i(z_0 - e) = 0)$

Identification

- When are α_i and or β_i continuous in Z_i near z_0?
- Imprecise control:
  - Suppose α_i and β_i depend on variables U_i and Z_i (and hence D_i) depends on U_i and V_i.
  - This allows for endogeneity of D_i because V_i may be correlated with U_i.
  - If the density of V_i conditional on U_i is continuous (imprecise control) and α_i and β_i are continuous functions of U_i then the continuity assumptions (local randomization) are satisfied.
  - This condition means that Z_i will not be a deterministic function of unobservables in the outcome equation – individuals do not have *precise* control over Z_i (and hence D_i).

Identification

- Regression kink.
  - The regression kink design estimand is:

$$\frac{\lim_{z \to z_0^+} \frac{\partial E(Y|Z=z)}{\partial z} - \lim_{z \to z_0^-} \frac{\partial E(Y|Z=z)}{\partial z}}{\lim_{z \to z_0^+} \frac{\partial E(D|Z=z)}{\partial z} - \lim_{z \to z_0^-} \frac{\partial E(D|Z=z)}{\partial z}}$$

  - A straightforward example of when this works is when $Y_1 - Y_0$ is constant, $E(Y_0 \mid Z = z)$ is continuous and has no kink at z_0, and $Pr(D = 1 \mid Z = z)$ is continuous but has a kink at z_0.
  - See Card, Lee, Pei, and Weber (2015) for more on a generalized kink design.

Identification

- When D_i is not binary.
  - If $E(D_i \mid Z_i = z)$ experiences a discontinuity at z_0 then we can use the same fuzzy design methods as above. The identification argument is essentially the same.
- With multiple thresholds,
  - we can use separate indicators for the thresholds as instruments in a 2SLS estimator. We can use RDD as a way of interpreting and presenting these results and as a way of considering the validity (identification).

Implementation

- Graphical evidence is important in these papers.

Implementation

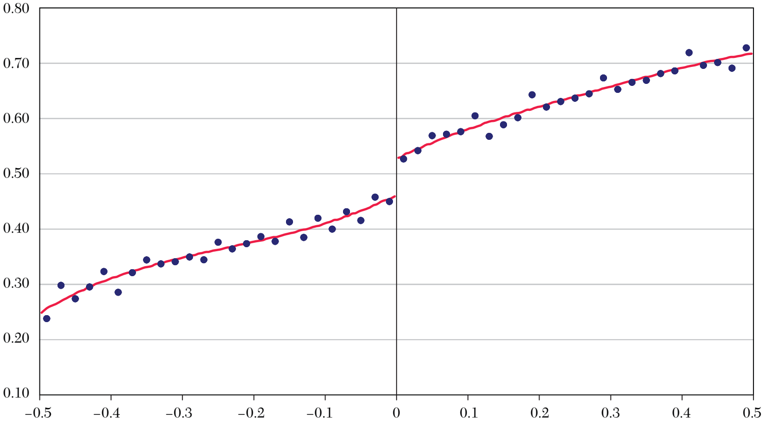


Figure 6. Share of Vote in Next Election, Bandwidth of 0.02 (50 bins)

Implementation

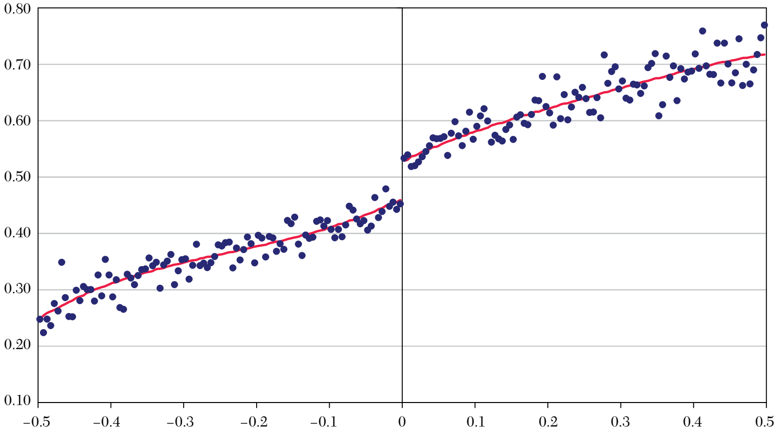


Figure 8. Share of Vote in Next Election, Bandwidth of 0.005 (200 bins)

Implementation

- Typical choice is to choice *B* bins and plot means within each bin.
- These bins should be smaller than the optimal bandwidth – this demonstrates the variability in the data.
- Calonico, Cattaneo, and Titiunik (2015) offer a data-driven approach.

Implementation

- Justifying the assumption of “imprecise control”:
  - Showing that covariates don’t exhibit a discontinuity – do an RDD with the covariate as the dependent variable.
  - The “forcing variable”, Z, should not exhibit a spike in its density at z_0 – McCrary test
  - Placebo tests –
    - in a study of the effect of retirement on mortality Fitzpatrick and Moore (2016): look at every other age cutoff and plot distribution of estimates

Implementation

- Using covariates:
  - test for discontinuity – use an aggregate statistic if there are many covariates (multiple testing problem)
  - include the covariates in the RD regression to improve efficiency
  - use them to residualize Y_i before doing RD

Robustness

- It is typical and recommended that an RD design includes
  - demonstration that slight decreases in bandwidth lead to expected results (corresponding to reduced bias but increased variance)
  - demonstration that a higher order polynomial leads to expected results (corresponding to reduced bias but increased variance)
  - demonstration that results are robust to different kernels, adding covariates, etc.
  - placebo tests

Identification

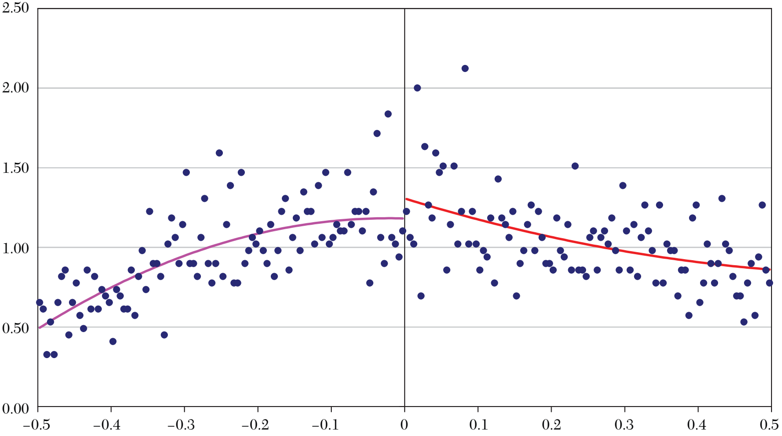


Figure 16. Density of the Forcing Variable (Vote Share in Previous Election)

Further reading

- Imbens and Lemiuex (2008)
- Hahn, Todd, and van der Klauww (2001)
- Card, Lee, Pei, and Weber (2015)
- Mattias Cattaneo at Michigan

Gauriot and Page (2015)

- In a 2015 article in *AER*, Gauriot and Page study individual-specific incentives in the game of cricket.
- The individual player (batsman) has incentives that misalign with the incentives of the team discontinuously when he has the opportunity to pass a symbolic landmark (50 or 100 or 200 runs in an innings).
- The strike rate (runs per ball) should jump discontinuously when the player's number of runs is just under these thresholds.
- This is a sharp RDD.

Gauriot and Page (2015)

- In a 2015 article in *AER*, Gauriot and Page study individual-specific incentives in the game of cricket.
- The individual player (batsman) has incentives that misalign with the incentives of the team discontinuously when he has the opportunity to pass a symbolic landmark (50 or 100 or 200 runs in an innings).
- The strike rate (runs per ball) should jump discontinuously when the player's number of runs is just under these thresholds.
- This is a sharp RDD.
- I know very little about cricket, so bear with me here!

Gauriot and Page (2015)

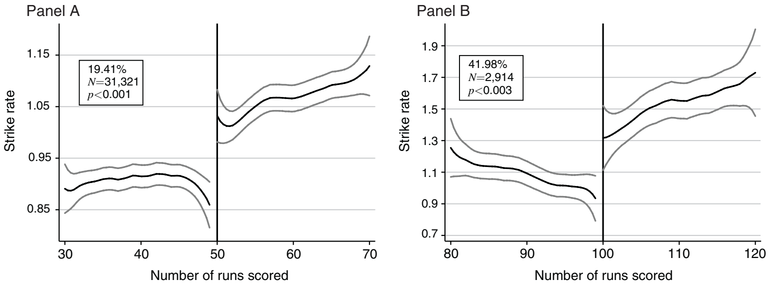


FIGURE 2. DISCONTINUITY IN THE STRIKE RATE (*Average number of runs per ball*) AROUND LANDMARKS 50 (*Panel A*) AND 100 (*Panel B*).

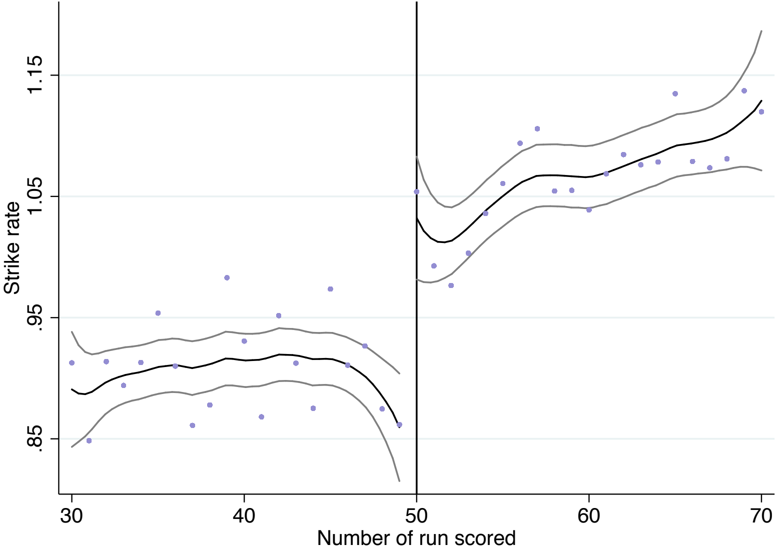
Notes: Local linear regression, triangular kernel, bandwidth of five runs. Panel A restricts the sample to players who reached 70 in the innings and panel B restricts the sample to players who reached 120 in the innings.

Source: ODI matches over the period 2001–2014.

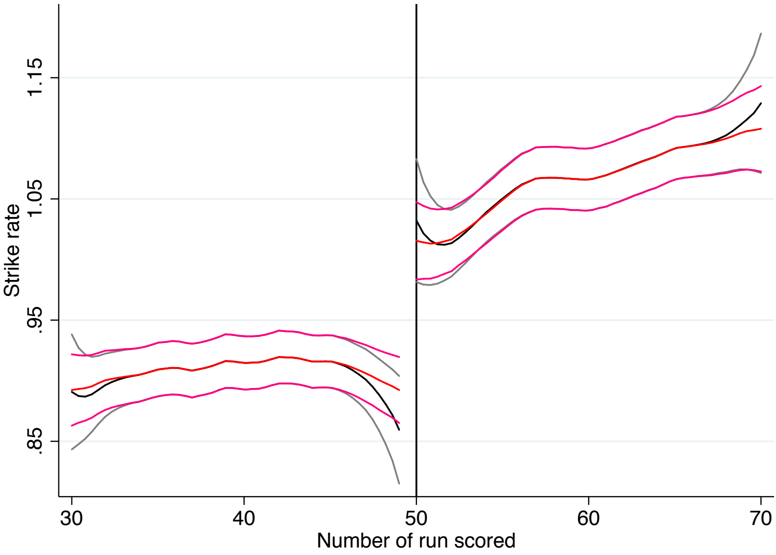
Gauriot and Page (2015)

- I will use this example to demonstrate:
  - different ways of estimating discontinuities
  - placebo test

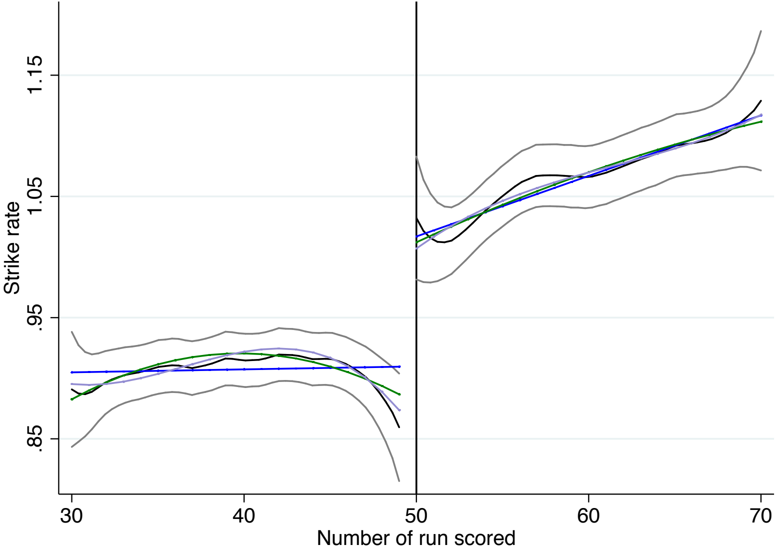
Gauriot and Page (2015)



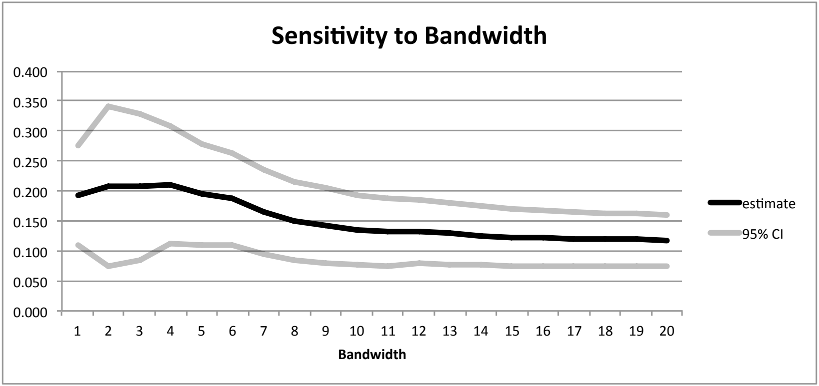
Gauriot and Page (2015)



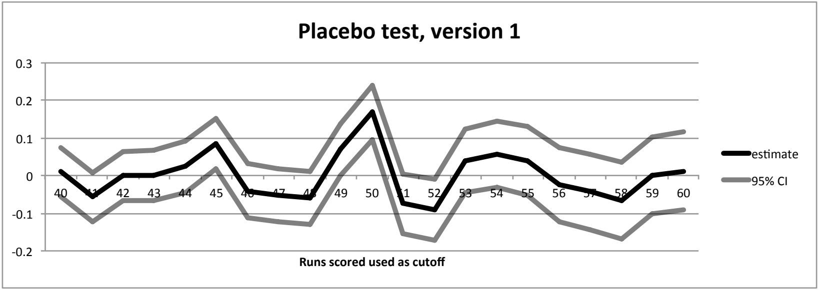
Gauriot and Page (2015)



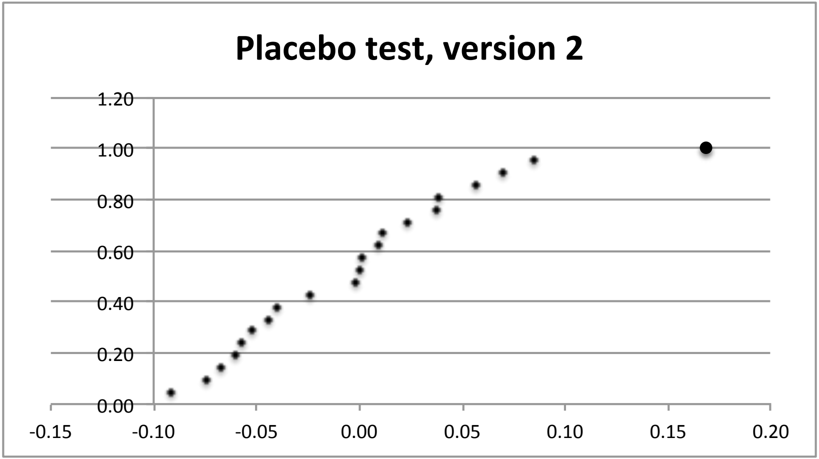
Gauriot and Page (2015)



Gauriot and Page (2015)



Gauriot and Page (2015)



Angrist et al. (2015)

- This paper compares RDD estimates from observational data to experimental estimates.
  - Reviewers scored scholarship applicants on a scale from 11 to 26.
  - If the score exceeded a certain threshold, an award was guaranteed.
  - If the score was below the threshold, the student was randomly assigned to either receive the award or not.
    - This was done within strata defined by colleges the student applied to.
  - The experimental estimate compares treatment and control for those just below the threshold.
  - The RDD estimate compares those just above the threshold to those just below it who did not receive an award (sharp RDD).

Angrist et al. (2015)

Table 2—Effects on Four-Year College Enrollment in Year Two

	Experimental sample (1)	Observational sample (2)	Experimental RD sample (3)	RD sample (4)
Control mean	0.639	0.708	0.685	0.685
Raw difference	0.142***	0.086***	0.107***	0.044
	(0.028)	(0.027)	(0.033)	(0.037)
Panel A. Strata-adjusted estimates				
Matching	0.144***	0.091***	0.116***	0.096***
	(0.024)	(0.023)	(0.027)	(0.031)
OLS	0.144***	0.091***	0.116***	0.099***
	(0.024)	(0.022)	(0.027)	(0.032)
Panel B. Esitmates with selection controls				
OLS	0.143***	0.094***	0.120***	0.107***
	(0.023)	(0.022)	(0.026)	(0.032)
OLS with r.v. controls				0.024
				(0.064)