Quantitative Policy Evaluation
Regression Discontinuity.

Eduardo Fé

Introduction: Lee (2001)

Let's start with a sobering thought

Introduction: Lee (2001)

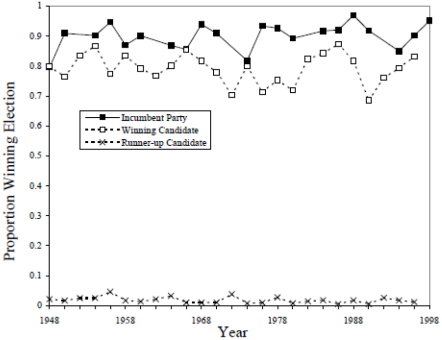
Let's start with a sobering thought

[...]*in-office parties have kept the White House two-thirds of the time when they have run incumbent candidates (R. Mayhew, 2008)*

Introduction: Lee (2001)

For the U.S. House of Representatives, in any given election year, re-election rate is about 80%

FIGURE I: Electoral Success of U.S. House Incumbents: 1948-1998



Note: Calculated from ICPSR study 7757. Details in Data Appendix. Incumbent party is the party that won the election in the preceding election in that congressional district. Due to re-districting on years that end with "2", there are no points on those years. Other series are the fraction of individual candidates in that year, who win an election in the following period, for both winners and runner-up candidates of that year.

Introduction: Lee (2001)

However stark data are, concluding that incumbency is responsible for this empirical regularity is not at all granted.

- Incumbents might be more charismatic.
- Incumbents might have more campaign resources.
- Voters might simply favour the winning party.
- Democrat (Republican) incumbents might be more successful simply because Democrat (Republican) incumbents tend to represent districts that are predominantly Democratic (Republican)!

Introduction: Lee (2001)

If we were to run an experiment to assess the effect of incumbency, what would we do?

We could run a completely randomized experiment:

- At time t, you could allocate election winners by flipping a coin,
- Then, congressional representatives would be comparable in traits, on average, across parties.
- The traits of districts with a Democrat/Republican representative would also be balanced on average.
- Now, move forward to the next election, at period $t+1$. Since incumbents were randomly selected at time t, any difference in re-election by incumbency would have a causal interpretation.

In practice we cannot run such experiment.

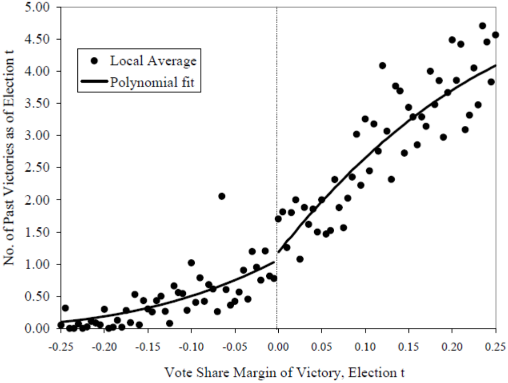
Introduction: Lee (2001)

Lee (2001) argues that, although the traits of winning and losing candidates are probably very different, it is likely that candidates who win a election by a slim margin are closely comparable to candidates who barely lose the election by a slim margin...

Introduction: Lee (2001)

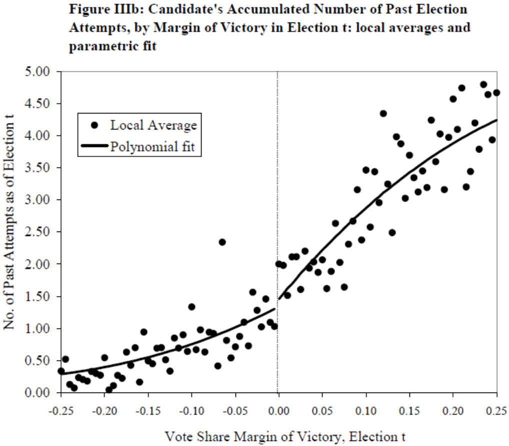
Indeed, in a neighbourhood of the 50% cut-off, Lee (2001) finds similarities in past political experience

Figure IIb: Candidate's Accumulated Number of Past Election Victories, by Margin of Victory in Election t: local averages and parametric fit



Introduction: Lee (2001)

and in electoral experience

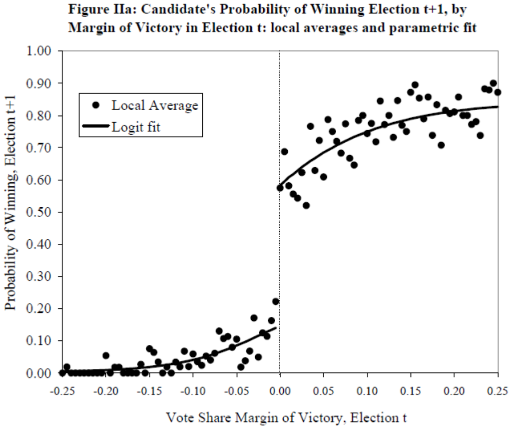


Introduction: Lee (2001)

So, it looks like around the 50% victory threshold, candidate's traits are comparable and, for a sufficiently small neighbourhood of the victory threshold (say, 2%) one could make the case that victory at t (and thus incumbency in the following election) is allocated almost *as if* at random...

Introduction: Lee (2001)

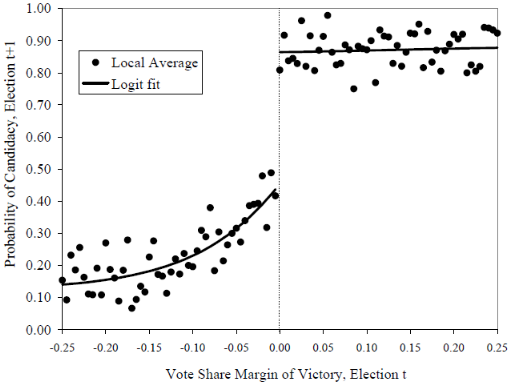
... in which case, incumbency looks a bit like a superpower closely fought re-election...



Introduction: Lee (2001)

... as well as future candidacy

Figure IIIa: Candidate's Probability of Candidacy in Election t+1, by Margin of Victory in Election t: local averages and parametric fit



Introduction: Thistlethwaite and Campbell (1960).

Regression Discontinuity was introduced by Thistlethwaite and Campbell (1960).

They were investigating the effect of public recognition on reception of scholarships and career plans.

In their paper, students qualified for scholarships if they reached a qualifying score in a test.

Among the students who reached the qualifying score, those with scores beyond a level received a Certificate of Merit, whereas the rest just received letters of commendation.

Certificates of Merit carried great public recognition: recipients' names were published in booklets distributed to colleges and universities and these students received 2.5 times more newspaper coverage.

Introduction: Thistlethwaite and Campbell (1960).

Thistlethwaite and Campbell (1960) persuasively argue that the group of commended students who narrowly missed receiving the Certificate of Merit would be comparable (in a randomized experiment sense) to the group of students who narrowly earned the Certificate of Merit.

A comparison of outcomes across these two groups might then have a causal interpretation.

Introduction: Thistlethwaite and Campbell (1960).

Indeed, Thistlethwaite and Campbell (1960) find similarities by score in an alternative cognitive ability test

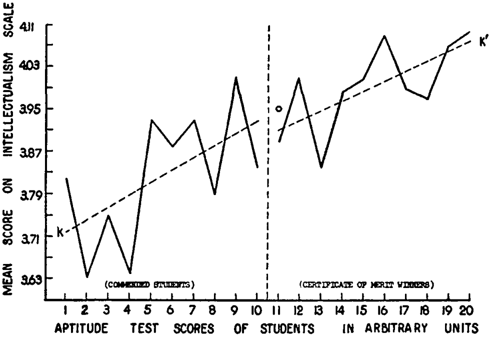


Fig. 4. Regression of attitudes toward intellectualism on exposure determiner.

Fig. 3. Regression of study and career plans on exposure determiner.

Introduction: Thistlethwaite and Campbell (1960).

Under their premise of quasi-randomization in a neighbourhood of the cut-off, the find that the Certificate of Merit lead to earning more scholarships

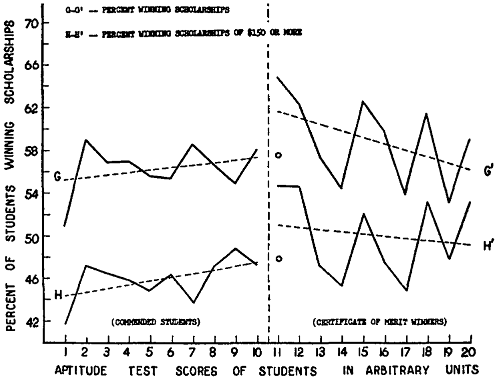


FIG. 2. Regression of success in winning scholarships on exposure determiner.

Introduction: Thistlethwaite and Campbell (1960).

Under their premise of quasi-randomization in a neighbourhood of the cut-off, the find that the Certificate of Merit lead to increased interest for doing a Ph.D. and M.D.

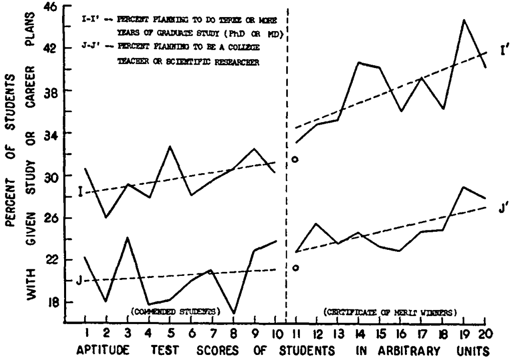


FIG. 3. Regression of study and career plans on exposure determiner.

Introduction: Happy trigger RD.

The idea of regression discontinuity took a while to be recognised; but since the publication of Hahn et al., 2001 the method got a life of its own...

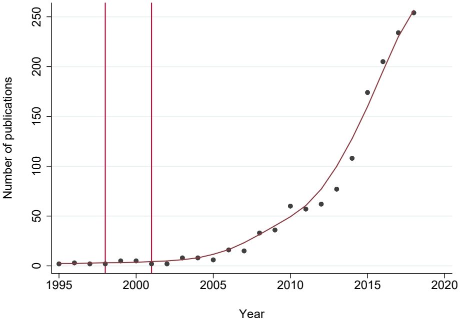


Figure: Number of articles published including the term 'Regression Discontinuity'. Results from a serch in Web of Science, June 2019. The red lines mark the date of Hahn, Todd and van der Klaauw's working paper and final publication in *Econometrica*.

Sharp Regression Discontinuity: Continuity-based framework.

We will approach Regression Discontinuity (RD) from two angles: a continuity-based framework (this session) and a randomization framework (a latter session). The continuity based framework departs slightly from the potential outcomes framework:

We have data from $i = 1 \dots N$ units, (Y_i, X_i, T_i), where

- Y_i is an outcome of interest
- X_i is a **continuous running variable** (e.g. the score in a test, a level of income, firm size, etc).
- T_i is a binary indicator taking value 1 if a unit receives an active treatment (otherwise $T_i = 0$). The peculiarity of the Sharp Regression Discontinuity design (SRD) is that a unit receives the active treatment if X_i exceeds a threshold/cut-off, say c. In other words, T is a deterministic function of X.

As before, $Y_i(t)$ denotes a unit's potential outcomes, only one of which can be observed (the selection problem is again operating here).

Sharp Regression Discontinuity: Continuity-based framework.

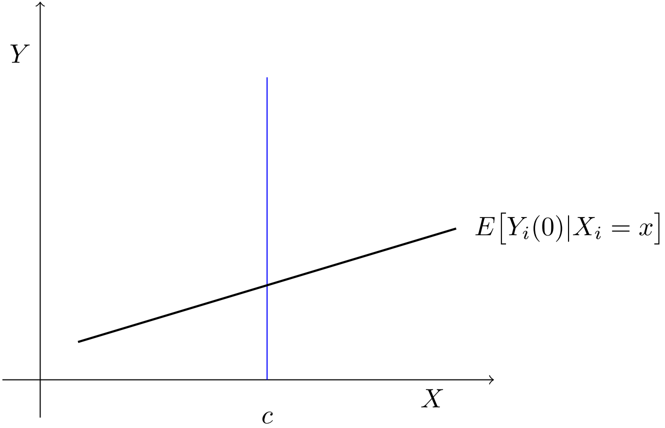
The idea behind SRD is to compare the outcomes of treated units that are marginally above the cut-off with control units that are marginally below the cut-off in order to estimate a causal effect.

The rationale is that treated and control units with scores of X falling within a small neighbourhood around the cut-off are likely to have, on average, similar pre-treatment observable and *unobservable* characteristics -thus these units are comparable.

Therefore, the any difference in average outcomes across treated and control groups in this neighbourhood must be due to the treatment only (which is the only factor distinguishing both groups).

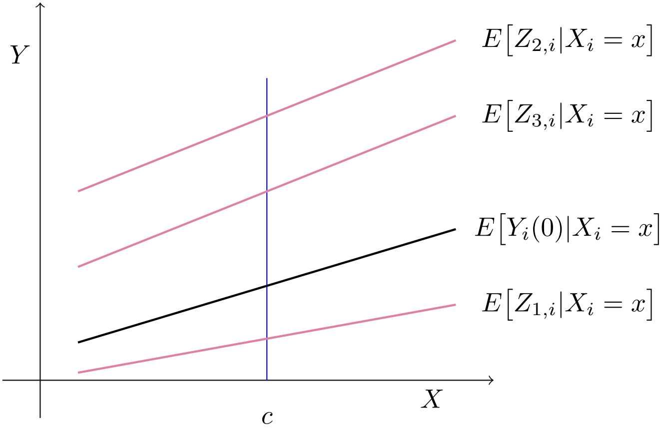
Sharp Regression Discontinuity: Continuity-based framework.

The underlying assumption is that since X is continuous, in the absence of treatment, Y should change smoothly along the range of X:



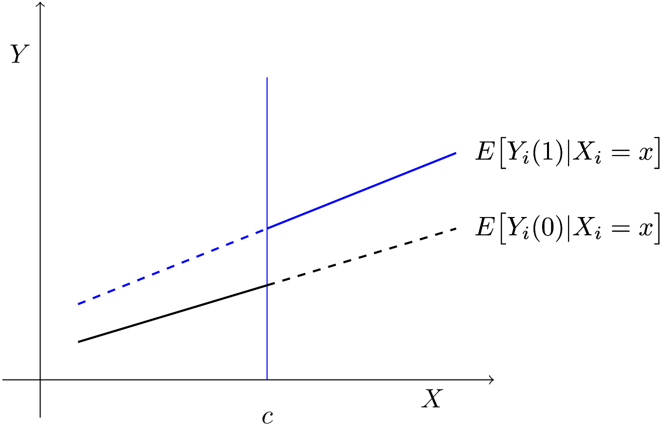
Sharp Regression Discontinuity: Continuity-based framework.

It is also implicit in the continuity-based framework that any pre-treatment variable, say Z, is continuous in X (and $Z(T) = Z$)



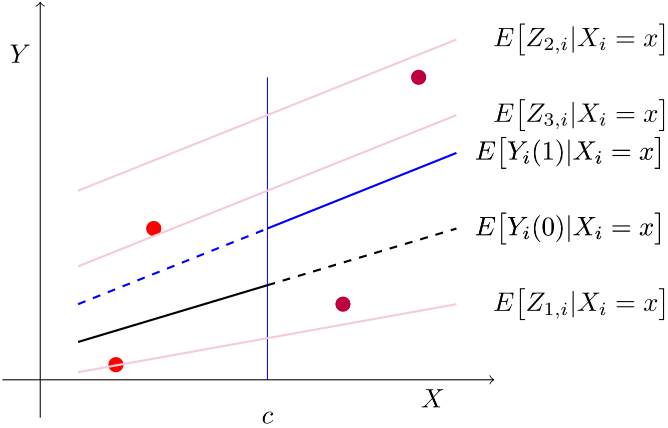
Sharp Regression Discontinuity: Continuity-based framework.

Now, if there is a treatment introduced when $X \geq c$, $T_i = 1$ and this treatment has a non-trivial effect, then for $X \geq c$ $Y_i(0)$ is not observable; for $X < c$ $Y_i(1)$ is not observable.



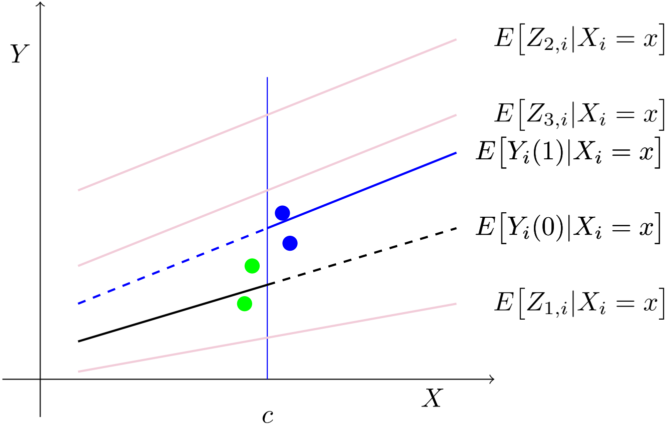
Sharp Regression Discontinuity: Continuity-based framework.

The purple and red units are not comparable -they are too different in their traits, Z_1, Z_2, Z_3, etc



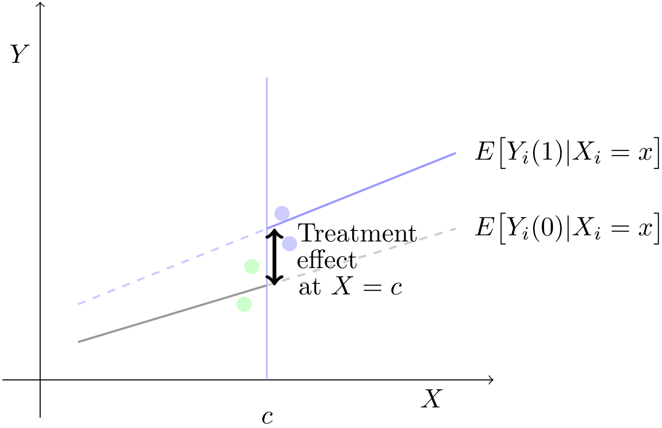
Sharp Regression Discontinuity: Continuity-based framework.

However, the blue and green units are comparable -their traits are not too different.



Sharp Regression Discontinuity: Continuity-based framework.

So the blue and green units can hopefully help us to estimate the treatment effect at $X = c$.



Sharp Regression Discontinuity: Continuity-based framework.

Overall, in the Sharp RD continuity-based framework the **conditional expectations** of the potential outcomes are continuous functions of x,

$$\mu_1(x) = E[Y_i(1)|X_i = x] \text{ and } \mu_0(x) = E[Y_i(0)|X_i = x] \tag{1}$$

We are interested in the treatment effect $\underline{\text{at } X = c}$,

$$\tau(c) = E[Y_i(1)|X_i = c] - E[Y_i(0)|X_i = c] = \mu_1(c) - \mu_0(c) \tag{2}$$

bearing in mind that the *observed* conditional expectation

$$\mu(x) = E\big[Y_i|X_i = x\big] = \begin{cases} \mu_1(x) & \text{if } x \geq c \\ \mu_1(x) & \text{if } x < c \end{cases} \tag{3}$$

does not reveal $\mu_0(c)$ (the problem of selection again), and in practice we might even lack observations at $X = c$

Sharp Regression Discontinuity: Continuity-based framework.

Hahn et al., 2001 show that if the conditional expectations are continuous in x at the cutoff level $x = c$, then

$$\tau(c) = \lim_{x \downarrow c} E[Y_i(1)|X_i = x] - \lim_{x \uparrow c} E[Y_i(0)|X_i = x] \tag{4}$$

The limits of $E[Y_i(1)|X_i = x]$ and $E[Y_i(0)|X_i = x]$ as x nears c from above or below can be easily estimated using nonparametric regression estimators.

The only caveat is that, in practice one often has not units with value of the score equal to the cutoff; so estimation at $x = c$ will generally need to rely on a degree of extrapolation (with information from nearby units being used to approximate the upper and lower limits of $\mu_{.}(x) = E\big[Y_i|X_i = x]$ at $x = c$).

Sharp Regression Discontinuity: Continuity-based framework.

$\tau(c)$ is a *Local* treatment effect in two senses:

- First it is only the average effect of the treatment at a particular value of the score x.
- Second it is only an average effect for *compliers*, in that it reveals the effect of treatment on units who would take upon the treatment only if they cross the threshold c.
- The latter is similar in the spirit to our discussion about LATE in IV settings; the difference in that in a Sharp Regression Discontinuity, everybody is a complier.

As a local parameter the external validity of $\tau(c)$ is limited; we cannot extrapolate to a broader population (with scores of x away from c).

Sharp Regression Discontinuity: Estimation.

To estimate

$$\tau(c) = \lim_{x \downarrow c} E[Y_i(1)|X_i = x] - \lim_{x \uparrow c} E[Y_i(0)|X_i = x] \tag{5}$$

when X is continuous, we can use a local polynomial regression. Specifically,

- A local polynomial regression at $x = c$, using observations with $X_i \geq c$ provides an estimate, $\hat{\mu}^+$, of $E[Y_i(1)|X_i = c]$.
- Similarly, a local polynomial regression at $x = c$, using observations with $X_i < c$ provides an estimate, $\hat{\mu}^-$, of $E[Y_i(0)|X_i = c]$.

The SRD estimator of $\tau(c)$ is given by

$$\hat{\tau}(c) = \hat{\mu}^+ - \hat{\mu}^- \tag{6}$$

Sharp Regression Discontinuity: Estimation.

The choice of kernel is discussed in several sources and is not a critical decision, although specific choices of kernel have more desirable theoretical properties[1]

A common choice in practice is the uniform kernel. When estimating $\hat{\mu}^-$ this kernel is defined as,

$$K(c;h) = \begin{cases} 1 & \text{if } c-h \leq X_i < c \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Essentially, this choice of kernel results in a least squares regression using only observations with X falling within $(c-h, c+h)$

[1]See Hardle, 1990; Li and Racine, 2006 for general discussion. For a discussion specific to RD settings, see Imbens and Lemieux, 2008, Cattaneo et al. or Fé, 2014

Sharp Regression Discontinuity: Estimation.

The choice of the order of the polynomial is more consequential.

In practice it is recommended to select $p = 1$ (local linear regression).

Estimates based on $p = 0$ have poor performance in the boundaries, (were some of the estimators we will discuss in this course need to be computed).

For a given choice of bandwidth, increasing p tends to increase the accuracy of the approximation, but also increases the variability of the estimator.

Large choices of p may incur in over-fitting, which increases the variability of the estimates.

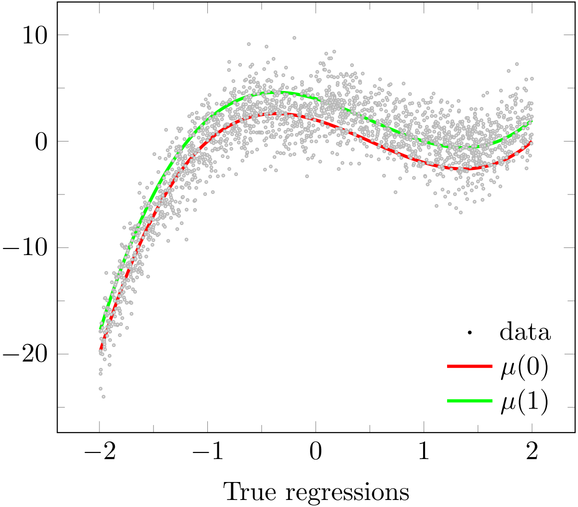
Sharp Regression Discontinuity: Estimation.

Bandwidth selection refers to choosing the neighbourhood around $x = c$ in which to estimate the effect $\tau(c)$. We consider data-driven methods only[2].

This is the most important decision in RD (Lee and Lemieux, 2010, Imbens and Lemieux, 2008, Imbens and Kalyanaraman, 2012, Calonico et al., 2014, Cattaneo et al., among others).

[2] Selection based on intuition or expert knowledge , though relatively common, lacks objectivity and encourages selection based on spurious criteria, such as statistical significance of the estimated effect.

Sharp Regression Discontinuity: Choice of bandwidth.



Sharp Regression Discontinuity: Choice of bandwidth.

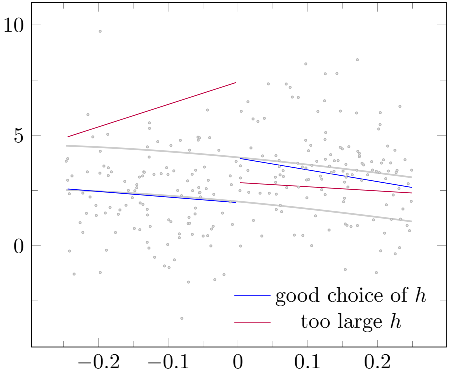


Figure: The grey lines are the true regression under treatment (top grey line) and control (lower grey line). The blue and red lines are local linear regressions at $x = 0$. The SRD estimate of the treatment effect equals to the gap between the red (blue) lines at $x = 0$. The results is very sensitive to the choice of bandwidth.

Sharp Regression Discontinuity: Estimation.

As in standard nonparametric regression,

- Smaller h will reduce the misspecification error but will tend to increase the variance of the estimated coefficients because fewer observations will be (effective) available for estimation.
- Larger h will result in more smoothing, less variability in the results but a bigger bias.
- Data-driven bandwidth selectors will normally try to balance this **bias-variance trade-off**.
- Although in practice it is possible to select a single bandwidth for estimation above and below the threshold, it is often useful to select different bandwidths, particularly when the density of data and its variability differs on each side.

Sharp Regression Discontinuity: Notes on Inference.

In principle, we can construct confidence intervals for $\tau(c)$ based on the asymptotic approximation,

$$CI = \left(\hat{\tau}_{SRD} \pm 1.96\sqrt{\hat{V}}\right) \tag{8}$$

for some estimator of the variance of $\hat{\tau}_{SRD}$, $\hat{V}$. Similarly tests of hypothesis could be based on the ratio

$$\frac{\hat{\tau}_{SRD} - \tau_{SRD}}{\sqrt{\hat{V}}} \tag{9}$$

However, neither of these methods are particularly well behaved in practice: they ignore the bias inherired from the nonparametric regression estimator, and they also ignore the variability due to having to estimate that bias.

Sharp Regression Discontinuity: Notes on Inference.

The best practice is to re-centre the confidence intervals and tests using an estimator of the bias, say $\hat{B}$ and replacing $\hat{V}$ with a robust version, say $\tilde{V}_b c$, that takes into account additional variation due to having to estimate the bias term. Therefore, recommended practice is to rely inference on

$$CI^* = \left((\hat{\tau}_{SRD} - \hat{B}) \pm 1.96\sqrt{\tilde{V}} \right) \tag{10}$$

for some estimator of the variance of $\hat{\tau}_{SRD}$, $\hat{V}$. Similarly tests of hypothesis could be based on the ratio

$$\frac{\hat{\tau}_{SRD} - \hat{B} - \tau_{SRD}}{\sqrt{\tilde{V}}} \tag{11}$$

Evaluating the quality of the design

Pre-test covariates One of the key premises underlying the SRD is that, in a neighbourhood of the cut-off, units should be comparable.

We can evaluate this question studying the continuity of variables that were determined before the treatment is assigned

These pre-test covariates should not be affected by the treatment, therefore a SRD analysis of these covariates at the cut-off should reveal no significant variation.

To undertake these tests, the methods discussed above are appropriate.

Evaluating the quality of the design

Placebo outcomes In some applications we might have access to additional outcomes that, not being pre-test covariates, should not be affected by the treatment.
We can apply a SRD analysis of these placebo outcomes at the cut-off, with the expectation that the analysis should reveal any significant variation around the cutoff.

Evaluating the quality of the design

The distribution of the running variable The number of observations around the cutoff should be reasonably similar

Large variation in the distribution of units at each side of the threshold might indicate non-random behaviour that might invalidate the SRD design. For instance, some units might be able to systematically manipulate their score to fall above/below a give threshold.

Said different, if units do not have the ability to manipulate their scores of the running variable, then the distribution of observations should be uniform around the threshold.

Evaluating the quality of the design

Other checks

- Sensitivity of bandwidth choice
- Sensitivity to observations close to the threshold. If no manipulation of the score are taking place, then results should not vary much if we drop observations with $|X_i - c| \leq \delta$ for a small $\delta > 0$

Local randomization inference approach

The continuity of potential outcomes in X is troublesome in practice.

This is particularly true when the running variable is time to an event. In these cases, time is often measured in months or years.

Not only is continuity violated in these cases, but doubts also emerge proximity to the cut-off. It seems plausible to argue that those who got a final mark of 69 in a exam are likely comparable in knowledge to those go to a 70 (though the latter would walk away with a distinction and the former with a 2:1)

However, those who are one year away from retirement and those who have just retired... are they comparable?

Local randomization inference approach

In a very important paper, Cattaneo et al., 2015 introduce a Fisherian randomization framework for Regression Discontinuity which does not rely on the assumption of continuity of $E(Y_i(t)|X)$.

The latter is replaced by assumptions similar to those explored in Fisher's Randomization Inference.

Local randomization inference approach

The intuition is that, if units have no knowledge of the cutoff or cannot manipulate their score of the running variable, then units that are close to the cut-off have the same chance of being barely above or barely below the cutoff.

If this is true, the Regression Discontinuity design might create quasi-unconfounded variation in the treatment.

Local randomization inference approach

Formally, Cattaneo et al. (2015) impose the following assumptions. Given a small neighbourhood around the cutoff:

Unconfounded assignment. The distribution of the running variable inside the neighbourhood does not depend on the unit's potential outcomes, is the same for all units, and is 'known'.

Exclusion restriction. The potential outcomes do not depend on the value of the running variable inside the neighbourhood. $Y_i(t, x) = Y_i(t)$ inside the neighbourhood.

Local randomization inference approach

REMARK: We are assuming beyond 'randomization' in the neighbourhood of the cutoff.

Specifically, randomization within the neighbourhood does not rule out that the potential outcomes are independent of x.

For instance, Cattaneo et al (forthcoming[3]) provide the following example

In a close election, even if victory was allocated at random, the winning party might attract more campaign funds in the future -as donors might believe that the voters might support the party again. If donations do increase chances of electoral victory, then there would be a positive relationship between the running variable and the potential outcomes (victory in the future)

[3] Cattaneo, M., R. Titiunik and G. Vazquez-Bare (forthcoming) Handbook of Research Methods in Political Science and International Relations, eds. L. Curini and R. J. Franzese, Sage Publications.

Local randomization inference approach

Under the unconfoundedness and exclusion assumptions, we can treat the regression discontinuity design as a randomized experiment near the cutoff.

Then, we can apply the techniques for Randomization Inference seen earlier in the course.
Specifically, let $x = c$ be the location of the cutoff. We can start by defining the *sharp null hypothesis*

$$H_0 : Y_i(1) = Y_i(0) \text{for all } i \text{ in a neighbourhood of } c \tag{12}$$

Then it is just a matter of implementing randomization inference in the neighbourhood.

Local randomization inference approach

How to choose the neighbourhood?

The method proposed by Cattaneo et al. (2015) is based on the idea that in a randomized experiment units' observed characteristics ought to be balanced across the treatment and control groups.

Therefore, the approach to finding the neighbourhood is to find a neighbourhood were we cannot reject that the pre-determined characteristics of treated and control units are no average identical.

Starting with a small neighbourhood, one conducts a test of the null hypothesis that the pre-treatment variables are balanced; if the null is not rejected, we enlarge the neighbourhood and retest... We select the largest neighbourhood in which the null hypothesis is not rejected.

Local randomization inference approach

As usual, we cannot directly test the assumptions of this approach. To defend the design in practice we can do a number of things

Density of the running variable. Under the assumptions, we should expect approximately the same number of units on each side of the cutoff.

Alternative cutoff values. No treatment effect should be foundat artificial cutoffs, since the treatment is not changing.

Robustness of small changes in the size of the neighbourhood

Fuzzy Regression Discontinuity

The SRD assumes that compliance with treatment is perfect

In practice, we often encounter situations when that is not the case and units that cross the cutoff do not take the treatment, while some units that do not cross the cutoff take the treatment

The situation is, essentially, identical to IV... but IV in a neighbourhood of the cutoff!

Fuzzy Regression Discontinuity

Let $Z_i = \mathbf{1}_{X_i \geq x}$ equal 1 if a unit is assigned to treatment (that is, if the unit crossed the cutoff); otherwise $Z_i = 0$

Let $T_i = 1$ if a unit actually took the treatment.

Remember that LATE, in the standard instrumental variable setting was defined as

$$\text{LATE} = \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{P(T_i|Z_i = 1) - P(T_i|Z_i = 0)} \ldots \tag{13}$$

Fuzzy Regression Discontinuity

In a Regression Discontinuity setting the we will focus on

$$\text{LATE}(c) = \frac{\lim_{x \downarrow c} E(Y_i|X_i = x) - \lim_{x \uparrow c} E(Y_i|X_i = x)}{\lim_{x \downarrow c} P(T_i|X_i = x) - \lim_{x \uparrow c} P(T_i|X_i = x)} \tag{14}$$

where each term can be estimated by a nonparametric regression at $x = c$ using data just above/below the cutoff.

Bandwidth selection is discussed in Imbens and Kalyanaraman, 2012.

References I

Calonico, S., **Cattaneo, M.**, and **Titiunik, R.** (2014). Robust nonparametric confidence intervals for Regression Discontinuity Designs. *Econometrica*, 82: 2295 2326

Cattaneo, M., **Frandsen, B.**, and **Titiunik, R.** (2015). Randomization inference in the Regression Discontinuity Design: An application to party advantages in the U.S. Senate. *Journal of Causal Inference*, 3: 1–24

Cattaneo, M., **Idobro, N.**, and **Titiunik, R.** (). A practical introduction to regression discontinuity designs: Foundations. To appear

Fé, E. (2014). Estimation and inference in regression discontinuity designs with asymmetric kernels. *Journal of Applied Statistics*, 41(11): 2406–2417

Hahn, J., **Todd, P.**, and **van der Klaauw, W.** (2001). Identification and estimation of treatment effects with a Regression Discontinuity Design. *Econometrica*, 69: 201–209

References II

Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press

Imbens, G.W. and **Kalyanaraman, K.** (2012). Optimal bandwidth choice for the Regression Discontinuity estimator. *Review of Economic Studies*, 1–27

Imbens, G.W. and **Lemieux, T.** (2008). Regression Discontinuity Designs: A guide to practice. *Journal of Econometrics*, 142: 615–635

Lee, D. and **Lemieux, T.** (2010). Regression discontinuity desings in economics. *Journal of Economic Literature*, 48: 281–355

Lee, D.S. (2001). The electoral advantage to incumbency and voters' valuation of politicians' experience: A regression discontinuity analysis of elections to the U.S... Working Paper 8441, National Bureau of Economic Research. doi:10.3386/w8441

Li, Q. and **Racine, J.** (2006). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press

References III

R. Mayhew, D. (2008). Incumbency advantage in U.S. presidential elections: The historical record. *Political Science Quarterly*, 123: 201–228. doi:10.2307/20203009

Thistlethwaite, D. and **Campbell, D.** (1960). Regression-Discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51: 309–317