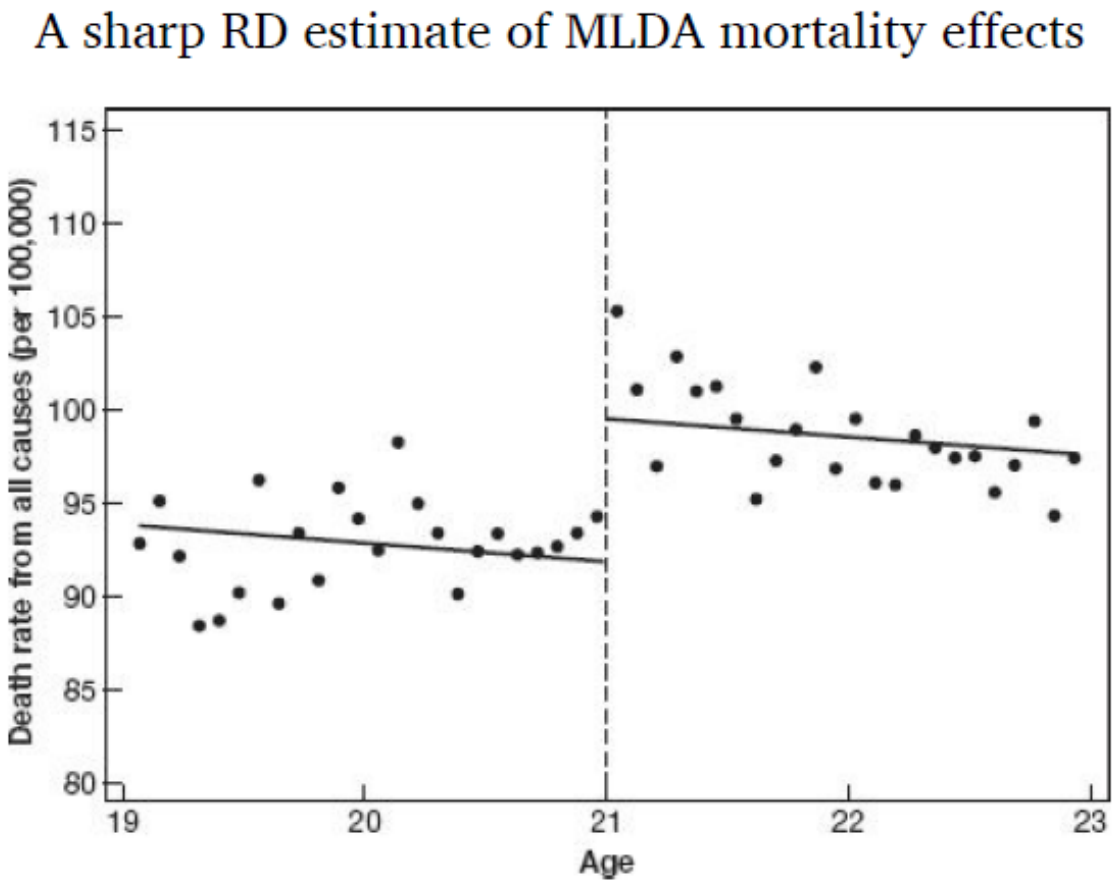
Regression Discontinuity

R Workshop

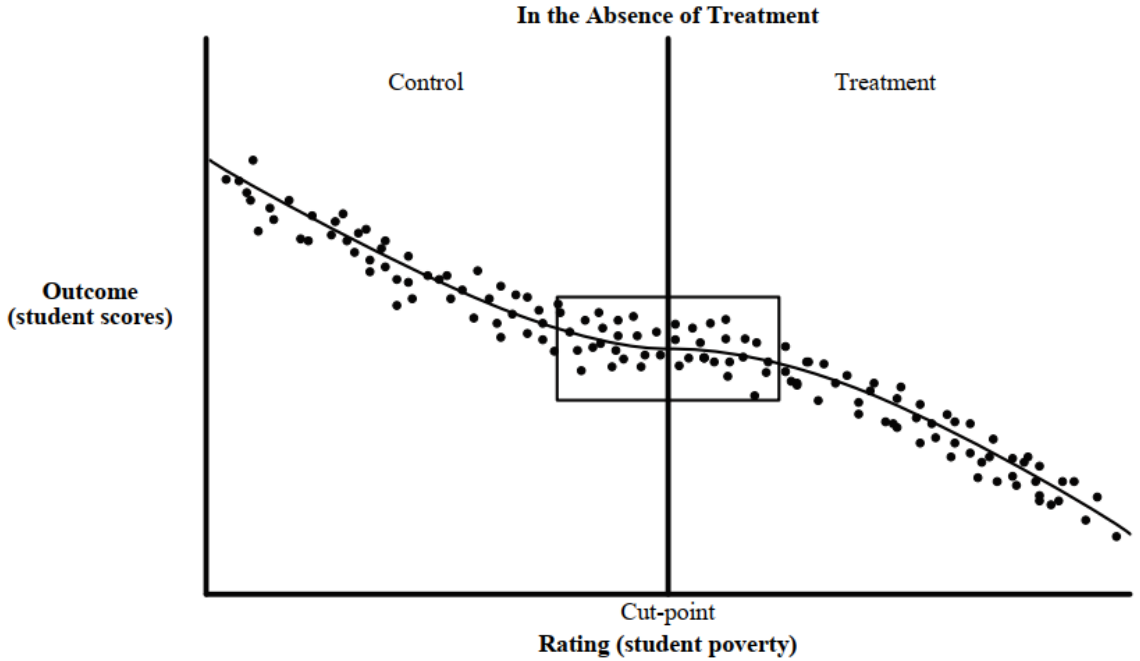
Intro to RD


A sharp RD estimate of MLDA mortality effects
Death rate from all causes (per 100,000)
115
110
105
100
95
90
85
80
19
20
21
22
23
Age

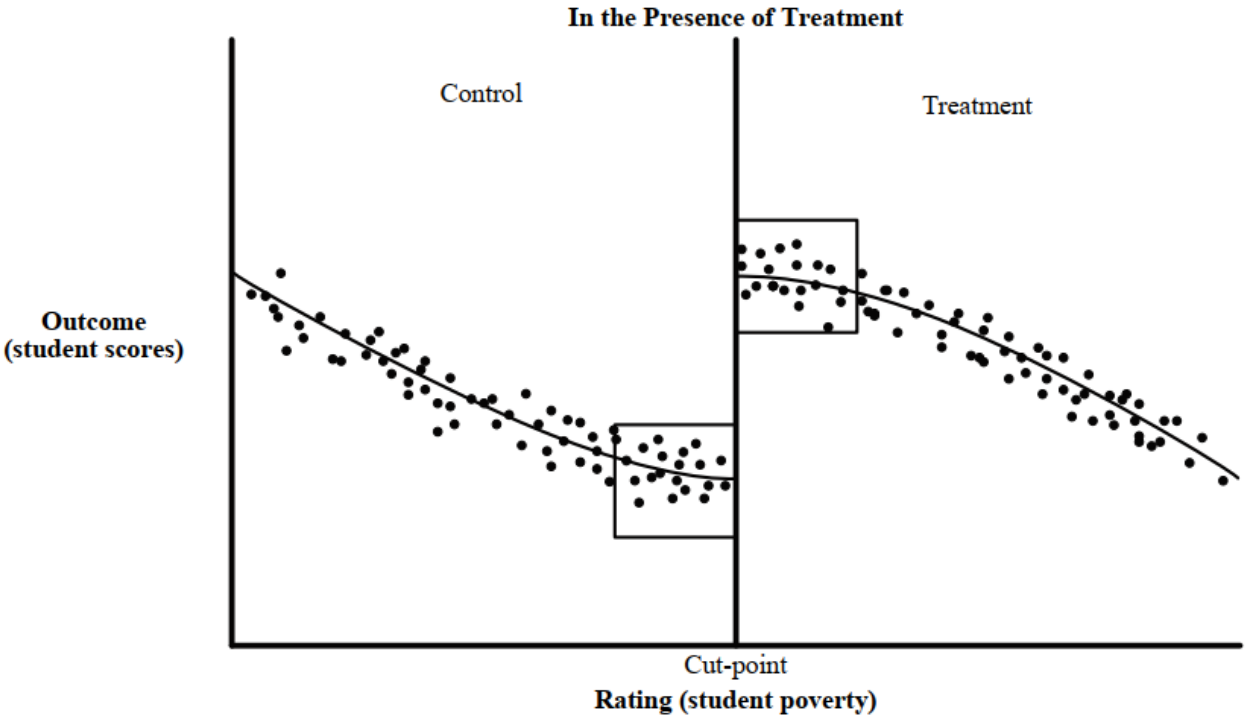
Intro to RD

Other examples: academic test scores, poverty levels, dates, elections...

Intro to RD


In the Absence of Treatment
Control
Treatment
Outcome
(student scores)
Cut-point
Rating (student poverty)

Intro to RD


In the Presence of Treatment
Control
Treatment
Outcome
(student scores)
Cut-point
Rating (student poverty)

The basic components

- The running/assignment variable: variable that determines whether you are in or out of the treatment.
- Cut-off point: specific value of that variable that determines in/out

The basic components

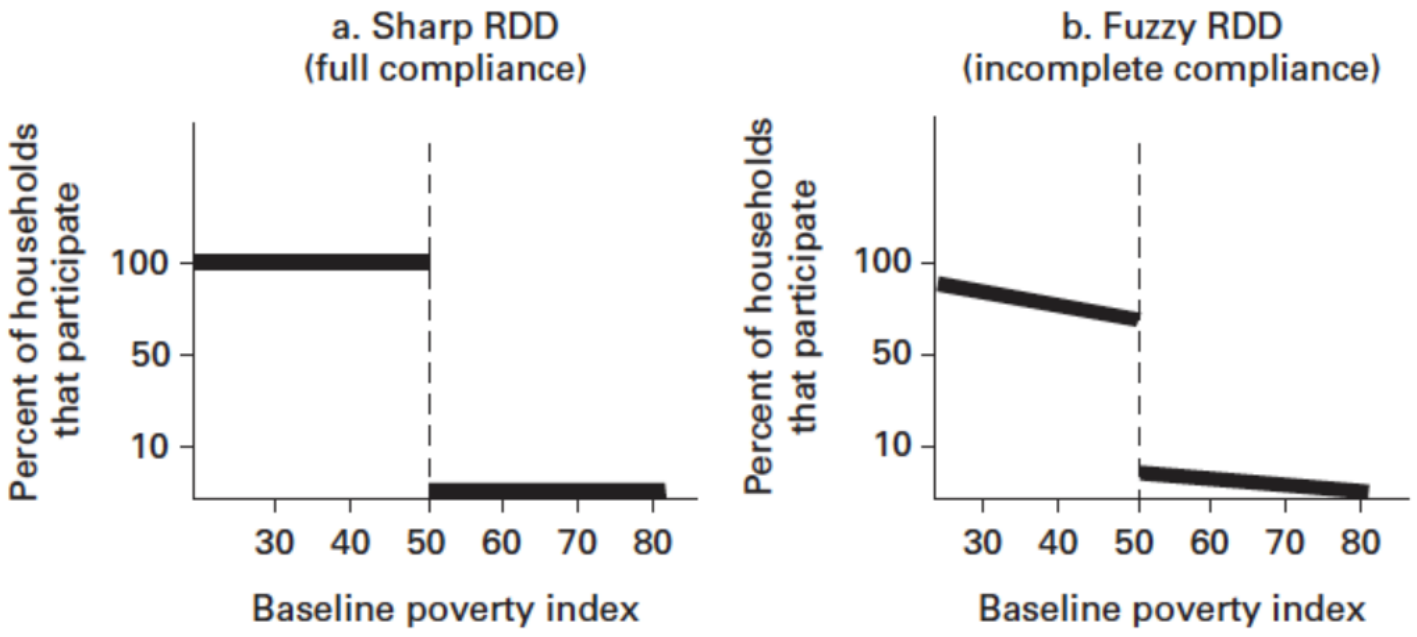
$$D_a = \begin{cases} 1 & \text{if } a \geq 21 \\ 0 & \text{if } a < 21 \end{cases}$$

Treatment = 1 if above cut-off

Treatment = 0 if below cut-off

Note: can be the other way around!

Fuzzy RD



World Bank

The basic model

Key assumptions:

1) only the treatment causes the discontinuity

2) the treatment is based on the cut-off value we observe

3) there is no discontinuity in other factors

X not randomly assigned, but whether X is right above or right below the cut-off point is essentially random

Estimation

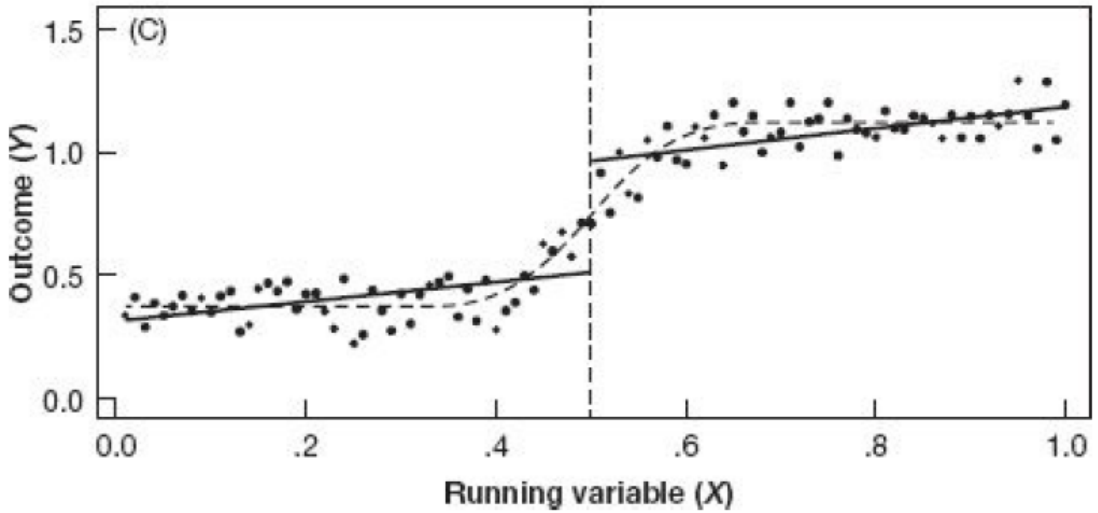
Two big ways of doing this: parametric and non-parametric.

Parametric: fitting the model using all the data in the sample.

Non-parametric: focusing only on the data points that are very close to the threshold.

Parametric: non-linearity

Consider this:



Parametric estimation: non-linearity

If the relationship isn't linear but we model it as such, we might confuse it with a jump.

To avoid this, we may need to add polynomials to our model.

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 \left(X_{1i} - C\right) + \beta_2 \left(X_{1i} - C\right)^2 + \epsilon_i$$

Parametric estimation: non-linearity

Polynomials can be sensitive and may produce larger bumps, so it is always a good idea to present both the linear model and the model with the polynomials.

Non-parametric estimation

If you sort of want to avoid all this mess, in some cases you may estimate your effect by only focusing on observations very near the threshold.

Here, the model specification won't really matter. A simple linear model is usually good enough.

How come?

Non-parametric estimation

Non-parametric estimation

Non-parametric estimation

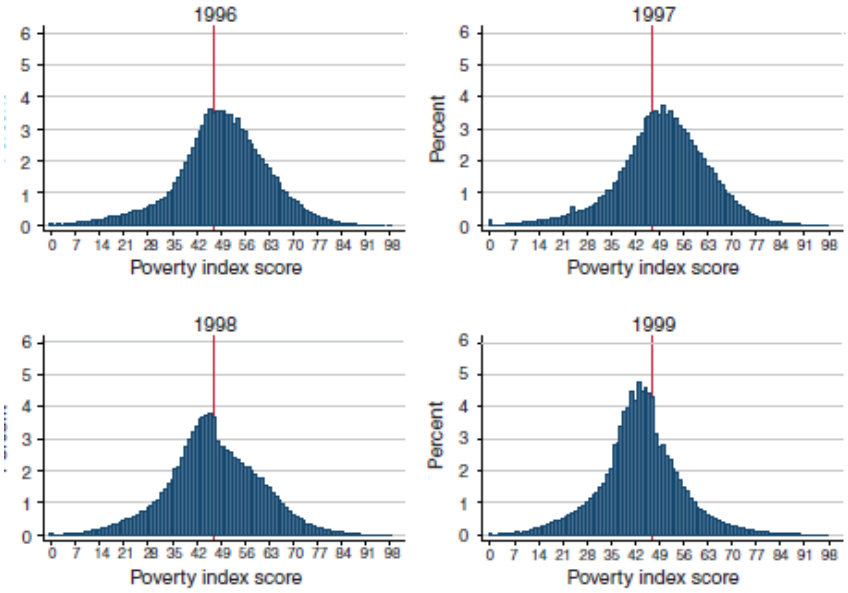
Big question here though is: how do you choose the size of that window?

There are some suggestions out there, and R has some packages that help with this.

But mainly it comes down to a **bias-variance** tradeoff (yes, again): The smaller your window, the less cases you will have (more variance), but the more accurate you will be (less bias).

Assumptions

The only thing happening at the threshold is the treatment



Assumptions

If cut-off point is known, there can be manipulation (from the part of people, or judges, or politicians, bureaucrats...). The point is that this ruins our assumptions that the people right above and right below the threshold are essentially the same.

Now we need to consider that those that have the treatment is because they have better contacts, more motivation, etc. So these people are no longer comparable.

Diagnostics

To check, the first thing we can use is qualitative work; look for evidence that this may or may not happen.

Second, graph the running variable (as shown above). If there is a clustering right above/below, this is suspicious.

Diagnostics

Third, we can make sure that other variables don't jump at the threshold (they shouldn't).

We can run a regression where we place these covariates as the outcome and make sure that the treatment has no significant effect.