# Week 5: Regression discontinuity designs

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2020

# Outline

- The logic of regression discontinuity
- Thistlethwaite and Campbell (1960)
- RDD and complete lack of overlap
- RDD and the potential outcomes framework
- Assumptions
- Examples
- Parametric estimation
- Fuzzy RDD and instrumental variables

# Basics

- We are going to back in time to the first application of RDD, although I'm going make some small changes so it's easier to explain (see the readings for the actual details)
- Thistlethwaite and Campbell (1960) studied the effects of winning a Certificate of Merit as part of the National Merit Scholarship
- The award was given to students who scored more than a given score $c$ (the cutoff point) in an aptitude test (SAT). If they scored $\geq c$, then they got the award
- $D = 1$ if $\geq c$ and zero otherwise. Let's say that the outcome $Y$ is income 5 years after receiving the award
- So the probability of treatment $P(D_i = 1|X_i)$, where $X_i$ is the score, jumps from 0 to 1 at the point $c$ (what did we call $P(D_i = 1|X_i)$?)

# Basics

- We could compare observed outcomes $E[Y_i|D_i = 1]$ and $E[Y_i|D_i = 0]$ but since we are on week 5 of the class you know that that's not going to get us far
- $E[Y_i|D_i = 0]$ is not a good prediction for $E[Y_{0i}|D_i = 1]$
- Treatment $D$ is not independent (nor mean independent) from the score $X$, and most aptitude test scores are also highly correlated with future income. Therefore, $X$ is a confounder, so at least we need to control for it
- Then we can compare $E[Y_i|D_i = 1, X_i]$ and $E[Y_i|D_i = 0, X_i]$ but now we have **complete lack of overlap** – we also most likely still don't have ignorability
- If observed, we could control for a vector of covariates $\boldsymbol{X}$, but we probably don't have all relevant confounders: intelligence, motivation, parent's education, quality of high school, and so on. All the factors that are associated with scoring well and future income

# RDD insight

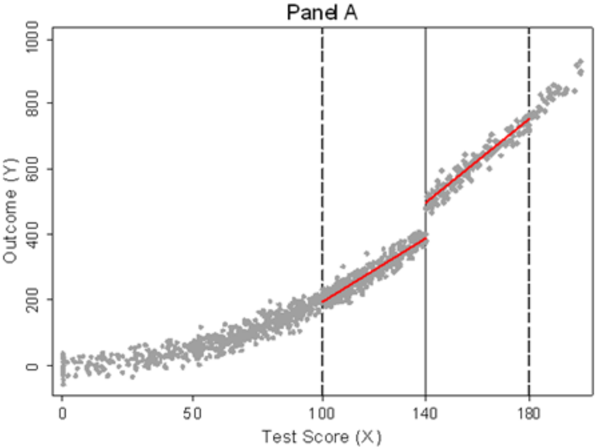
- Thistlethwaite and Campbell realized they could in fact compare a subsample of the students: **those who scored just above and just below of the cutoff** $c$
- By now I find this idea intuitive but it's not that intuitive at first. It helps if you notice that the point $c$ is arbitrary
- Is there a substantial difference between scoring, say, $c = 1200$ in the SAT versus 1210 or 1190? Probably not
- We know that the test scores are related to all sorts of observed and unobserved factors, but students scoring around 1,200 should be comparable
- Getting a score a little higher or a little lower than 1200 is random

# RDD insight

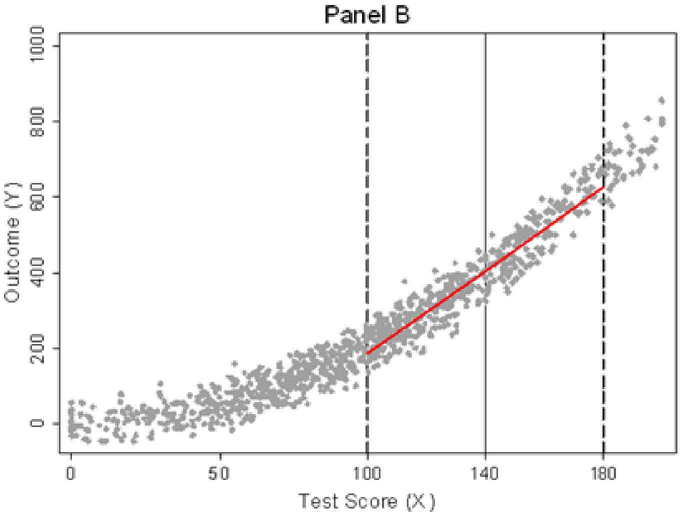
- If you follow this reasoning, then **close** to $c$ receiving the scholarship is as if it were at random
- You could image that –again, around $c$– an experiment was designed: choose an arbitrary cutoff point $c$ and give the award to those who scored a little after $c$ (caution, not the same as conditional randomization)
- Think at the limit to clarify. A student who scored 1000 is different from one that scored 1500
- A student who scored 1150 is different but not as much as a student who scored 1250
- A student who scored 1199.5 should be comparable to a student who scored 1200.5. One got the award, the other one didn't. Good luck, bad luck, respectively

# Example of positive treatment effect

- Simulated data with $c = 140$ and window (100, 180)



# No effect



Panel B
Outcome (Y)
0
200
400
600
800
1000
Test Score (X)
0
50
100
140
180

# Notation

- $X$ is the variable used to assign treatment (the SAT score in this example), which in RDD is called the **assignment variable** or the **running variable** or the **forcing variable** (note I didn't write $\boldsymbol{X}$, a vector of covariates)
- The reasoning above suggests that a comparison of $lim_{x \rightarrow c} E[Yi|Xi = x]$ and $lim_{x \leftarrow c} E[Yi|Xi = x]$ would provide an estimate of treatment effects (note the direction of the arrows)
- The above is equivalent to: $lim_{x \rightarrow c} E[Yi|Xi = x, D_i = 0]$ and $lim_{x \leftarrow c} E[Yi|Xi = x, D_i = 1]$ since in this example to right the right of $c$ everybody gets treatment; to the left nobody does
- Essentially, what we are saying is:

  $lim_{x \rightarrow c} E[Yi|Xi = x, D_i = 0] \approx E[Y_{0i}|X_i = c, D_i = 1]$ and

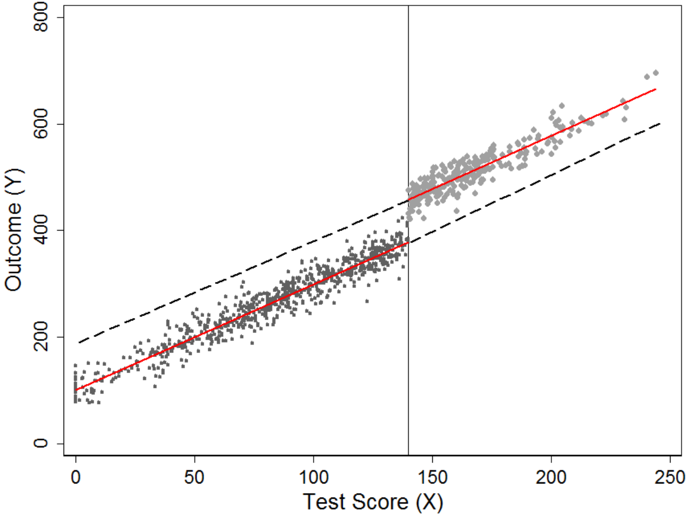
  $lim_{x \leftarrow c} E[Yi|Xi = x, D_i = 1] \approx E[Y_{1i}|X_i = c, D_i = 0]$
- Again, we don't really need to condition on $D$ but it makes easier to explain (I think)

# No overlap

- In the class on propensity scores we saw that the definition of overlap is for all $\boldsymbol{X}_i \in \varphi$, where $\varphi$ is the support (domain) of the covariates $\boldsymbol{X}_i$, $0 < P(D = 1|\boldsymbol{X}_i) < 1$
- We only have one variable $X$ here. In RDD (at least, **sharp RDD**) there is no overlap at all: $P(D = 1|X_i) = 0$ if $X < c$, while $P(D = 1|X_i) = 1$ if $X \geq c$
- Treatment effects around $c$ are based on **pure extrapolations**
- So we already know the main difficulty in estimating causal effects in RDD: we need to model correctly the relationship between $Y$ and the running variable $X$ or we will get our counterfactual predictions wrong
- Or said another way, **our model must be correctly specified**

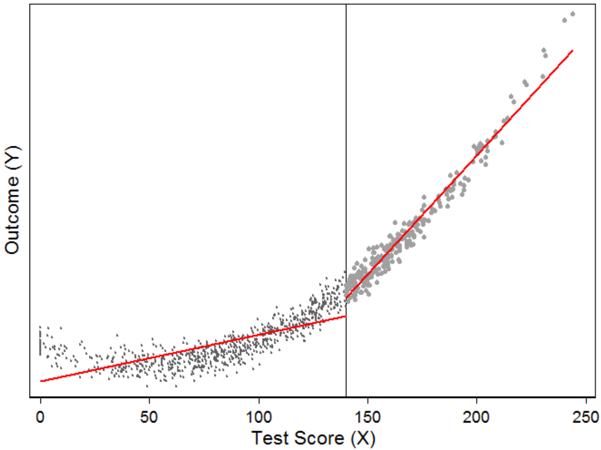
# Extrapolation

Dashed lines are extrapolations. If real-life example were like this, life would be easier: perfect linear relationship, so extrapolation is not a problem



# A bit more realistic

- True relationship is non-linear, but we use a linear model and incorrectly find a positive treatment effect



# Key issues

- I just said that we need to correctly specify the functional form between $Y$ and $X$. **This immediately suggests nonparametric or semiparametric models** – that's where RDD analysis has moved to in the last 5 to 10 years
- We will use both parametric and semiparametric models in the second part of the semester
- How close enough to $c$ is close enough? What window around $c$ should we use?
- It's a **bias-variance trade-off**: the closer to $c$ we are the better the assumptions hold, so we reduce bias. But closer to $c$ we use fewer and fewer observations, so the variance increases
- A larger window (bandwidth) means that we rely more on extrapolation
- We will see some "optimal" bandwidth approaches, but the usual approach is to perform sensitivity analyses using different bandwidths and checking balance

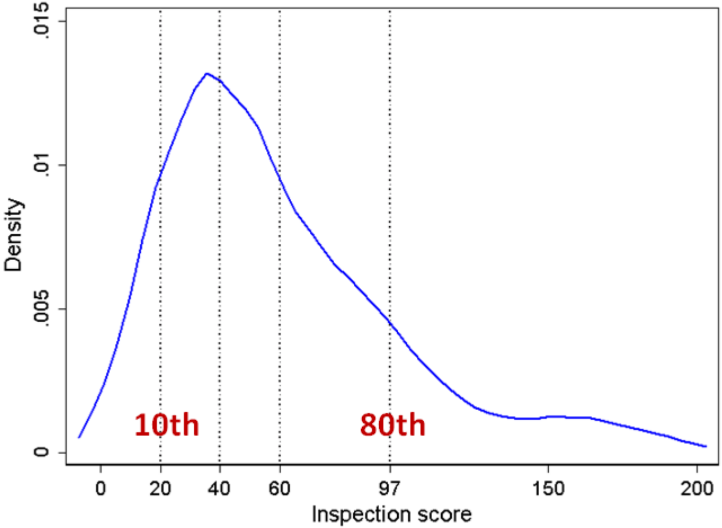
# Assumptions

- When would RDD not be valid? In econometrics, the usual condition is that there shouldn't be **manipulation with precision**
- If a student could know that she is close to 1200, the she could stop the test because she knows that she has the award already **Common confusion**: Some manipulation is fine (you can always study harder, for example). Manipulation **with precision** or the absence of a deterministic relationship between cut-off point and outcomes is key
- Think a different example: the running variable is a measure of carbon monoxide in blood. There could be a natural cutoff point $c$ that implies that after $c$ dead is imminent. Therefore, $c$ is not arbitrary when the outcome is death
- But with measurement error, then very close to $c$ it could be arbitrary
- Another example: 1500 as the cutoff point for very-low birth weight babies (VLBW)

# Examples

- Almond et al. (2010): Assignment variable is birth weight. Infants with low birth weight ($< 1,500$ grams or about 3 pounds) receive more medical treatment
- Lee, Moretti, Buttler (2004): The vote share (0 to 100 percent) for a candidate is a continuous variable. A candidate is elected if he or she obtains more than 50% of the votes. They evaluated voting record of candidates in close elections
- Perraillon at al. (2019): CMS rates nursing homes using 1 to 5 stars. Overall stars are assigned based on deficiency data transformed into a points system. Outcome: new admissions six months after the release of ratings (consumer response)

Assignment of stars based on scores



# RDD as a special type randomization

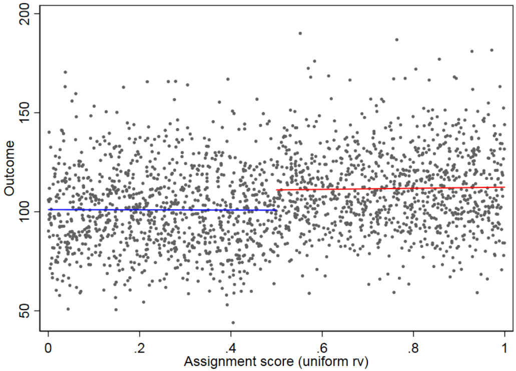
- Suppose you randomize people the old-fashioned way. You have a dataset with 2000 persons ids. You create a new column that is a draw from a uniform random variable called rv
- If $rv > 0.5$, then assign to treatment group. We know that each person has equal probability of being in either group (it's a uniform distribution)
- If no treatment is performed , would there be any relationship between an outcome –any outcome– and the uniform random variable?
- No. Furtheremore, there **wouldn't be any relationship between the assignment variable rv and any baseline characteristic** (rv and everything else are independent)
- But what about an outcome **after** performing an intervention on the treatment group? Is there a relationship between rv and the outcome? Said another way, do we need to control for rv in our models? No

# RDD as a special type randomization

```
set obs 2000
gen id = _n
* Simulated baseline outcome (chi-squared)
gen y0 = rnormal(10,1)^2
* Randomize
gen rv =uniform()
gen T=0
replace T=1 if rv >.5
* Pretend treatment is effective
gen     y1 = y0
replace y1 = y0+10 if T==1
```

# RDD as a special type of randomization

```
scatter y1 rv, msize(tiny) || lfit y1 rv if T==1, color(red) ///
        || lfit y1 rv if T==0, color(blue) ///
legend(off) ytitle("Outcome") xtitle("Assignment score (uniform rv)") ///
saving(rv.gph)
graph export rv.png, replace
```



# RDD as a special type of randomization

```
 * Controlling or not for the assignment variable is irrelevant
 * rv is not a confounder
qui reg y1 T
est sto m1
qui reg y1 T rv
est sto m2
est table m1 m2, p
----------------------------------------
    Variable |     m1           m2
-------------+--------------------------
           T |  10.78717    10.198423
             |    0.0000       0.0000
          rv |               1.1523433
             |                  0.7189
       _cons | 100.96664    100.68496
             |    0.0000       0.0000
----------------------------------------
                             legend: b/p
```

# RDD as a special type of randomization

- The difference is that in RDD the assignment variables is expected/assumed to be associated with $Y$
- So we have to control for that variable, unlike rv in the example above
- It does help to make analogies with randomization trials

# Parametric estimation

- Simplest case is linear relationship between $Y$ and $X$:
  $Y_i = \beta_0 + \beta_1 D_i + \beta_3 X_i + \epsilon_i$

  (Digression: Note how you can estimate the above regression without any hint of a problem even when there is no overlap at all)
- $D_i = 1$ if subject $i$ received treatment and $D_i = 0$ otherwise. We can also write this as $D_i = \mathbf{1}(X_i \geq c)$ or $D_i = \mathbb{1}_{[X_i \geq c]}$
- But we want $\beta_1$ at $c$, so we **center the running variable**:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_3 (X_i - c) + \epsilon_i$$

- Now the treatment effect is $\beta_1$ at $c$
- $E[Y_i|D_i = 1, X_i = c] = \beta_0 + \beta_1$ and $E[Y_i|D_i = 0, X_i = c] = \beta_0$, so:
  $E[Y_i|D_i = 1, X = c] - E[Y_i|D_i = 0, X_i = c] = \beta_1$
- We could add an interaction too. We could add quadratic terms to make the functional form more flexible at either side of the cutoff point. We will use nonparametric estimation in the second part of the semester

# Where are the covariates??

- If the assumptions of RDD hold, then all observed and unobserved covariates are balanced and we don't need to include them
- In practice we probably want to include the most relevant ones to improve precision, so the parametric model would be:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_3 (X_i - c) + \mathbf{X}_i' \beta + \epsilon_i$$

- We will see ways of incorporating covariates in nonparametric models

# Sharp and fuzzy RDD

- **Sharp RDD**: Assignment or running variable completely determines treatment. A jump in the probability of treatment before and after cutoff point, from 0 to 1
- **Fuzzy RDD**: Cutoff point increases the *probability* of treatment but doesn't completely determines treatment. A change in the probability of treatment before and after but not from 0 to 1 (i.e. there is some overlap now)
- Which brings us back to the world of instrumental variables
- Fuzzy RDD is not used as often but has a lot of potential in particular because no mental contortions are needed to justify the **IV exclusion restriction**
- Think of encouragement designs or imperfect compliance
- We will go over the basics of IV next class