Lecture 4: Regression Discontinuity Design

Fabian Waldinger

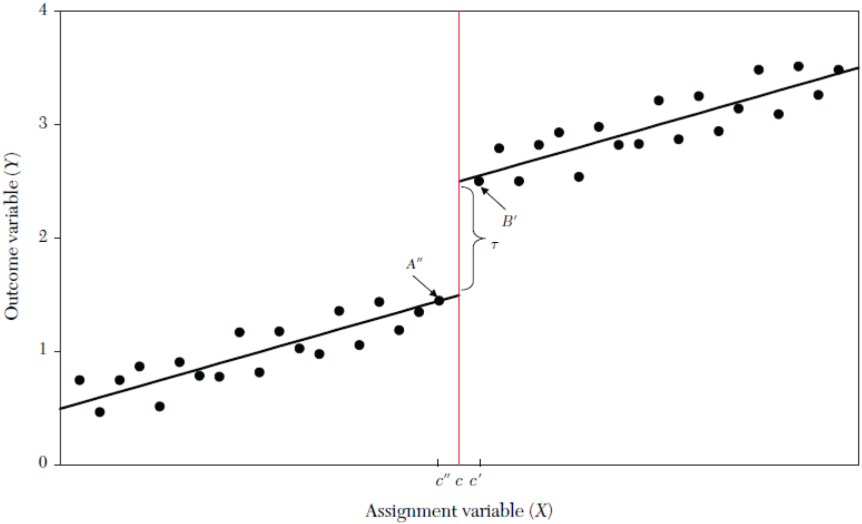
Topics Covered in Lecture

1. Sharp RD.
2. Fuzzy RD.
3. Practical Tips for Running RD Models.
4. Example Fuzzy RD: Angrist & Lavy (1999) - Maimonides Rule.
5. Regression Kink Design (very briefly).

Regression Discontinuity Design

- Regression discontinuity research designs exploit the fact that some rules are quite arbitrary and therefore provide good quasi-experiments when you compare people (or cities, firms, countries,...) who are just affected by the rule with people who are just not affected by the rule.
- There are 2 types of RD designs:
  1. Sharp RD: treatment is a deterministic function of a covariate X.
  2. Fuzzy RD: exploits discontinuities in the *probability* of treatment conditional on a covariate X (the discontinuity is then used as an IV).
- RD captures the causal effect by distinguishing the nonlinear and discontinuous function, $1(X_i \geq X_o)$ from the smooth function $f(X_i)$.

Example Linear RD



Sharp RD

- In Sharp RD designs you exploit that treatment status is a deterministic and discontinuous function of a covariate x_i.

  $\mathrm{D}_i = \{ \begin{matrix} 1 \text{ if } x_i \geq x_o \\ 0 \text{ if } x_i < x_o \end{matrix}$

  x_o is a known threshold or cutoff
- Once we know x_i we know D_i.
- As highlighted by Imbens and Lemieux (2008) there is no value of x_i at which you observe both treatment and control observations.
  $\rightarrow$ the method relies on extrapolation across covariate values.
- For this reason we cannot be agnostic about regression functional form in RD.

Formalizing Sharp RD

- Suppose that in addition to the assignment mechanism above, potential outcomes can be described by a linear, constant effects model:

$$E[\mathrm{Y}_{0i}|X_i] = \alpha + \beta X_i$$
$$\mathrm{Y}_{1i} = \mathrm{Y}_{0i} + \rho$$

- This leads to the regression:

$$\mathrm{Y}_i = \alpha + \beta X_i + \rho \mathrm{D}_i + \eta_i$$

- The key difference between this regression and regressions we have investigated in previous lectures is that D_i is not only correlated with X_i but it is a deterministic function of X_i.

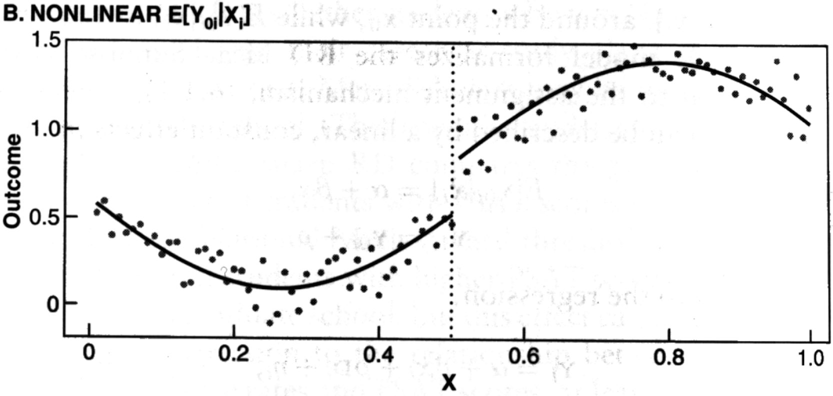
Key Identifying Assumption

- Key identifying assumption:

$$E[\mathrm{Y}_{0i}|X_i] \text{ and } E[\mathrm{Y}_{1i}|X_i] \text{ are continuos in } X_i \text{ at } X_0.$$

- This means that all other unobserved determinants of Y are continuously related to the running variable X.
- This allows us to use average outcomes of units just below the cutoff as a valid counterfactual for units right above the cutoff.
- This assumption cannot be directly tested. But there are some tests which give suggestive evidence whether the assumption is satisfied (see below).

- Sometimes the trend relation $E[Y_{0i}|x_i]$ is nonlinear.



Sharp Regression Discontinuity - Nonlinear Case

- Suppose the nonlinear relationship is $E[Y_{0i}|x_i] = f(X_i)$ for some reasonably smooth function $f(X_i)$.
- In that case we can construct RD estimates by fitting:

$$Y_i = f(x_i) + \rho D_i + \eta_i$$

- There are 2 ways of approximating $f(x_i)$:
  1. Use a nonparametric kernel method (see more below).
  2. Use a pth order polynomial: i.e. estimate:
     $$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p + \rho D_i + \eta_i \qquad (1)$$
- During the following introduction we will focus on the pth order polynomial approach but will discuss the other approach below.

Different Polynomials on the 2 Sides of the Discontinuity

- We can generalize the function $f(x_i)$ by allowing the x_i terms to differ on both sides of the threshold by including them both individually and interacting them with D_i.
- As Lee and Lemieux (2010) note, allowing different functions on both sides of the discontinuity should be the main results in an RD paper (as otherwise we use values from both sides of the cutoff the estimate the function on each side).
- In that case we have:

$$E[\mathrm{Y}_{0i}|X_i] = \alpha + \beta_{01}\widetilde{X}_i + \beta_{02}\widetilde{X}_i^2 + ... + \beta_{0p}\widetilde{X}_i^p$$
$$E[\mathrm{Y}_{1i}|X_i] = \alpha + \rho + \beta_{11}\widetilde{X}_i + \beta_{12}\widetilde{X}_i^2 + ... + \beta_{1p}\widetilde{X}_i^p$$

where $\widetilde{X}_i = X_i - X_0$

- Centering at X_0 ensures that the treatment effect at $X_i = X_0$ is the coefficient on D_i in a regression model with interaction terms (because you do not have to add values of the D_i interacted with X to get the treatment effect at X_0).

Different Polynomials on the 2 Sides of the Discontinuity

- To derive a regression model that can be used to estimate the causal effect we use the fact that D_i is a deterministic function of X_i :

$$E[\mathrm{Y}_i|X_i] = E[\mathrm{Y}_{0i}|X_i] + (E[\mathrm{Y}_{1i}|X_i] - E[\mathrm{Y}_{0i}|X_i])\mathrm{D}_i$$

- The regression model which you estimate is then:

$$\begin{aligned}\mathrm{Y}_i &= \alpha + \beta_{01}\widetilde{x}_i + \beta_{02}\widetilde{x}_i^2 + ... + \beta_{0p}\widetilde{x}_i^p \\ &+\rho\mathrm{D}_i + \beta_1^*\mathrm{D}_i\widetilde{x}_i + \beta_2^*\mathrm{D}_i\widetilde{x}_i^2 + ... + \beta_p^*\mathrm{D}_i\widetilde{x}_i^p + \eta_i \qquad (2)\end{aligned}$$

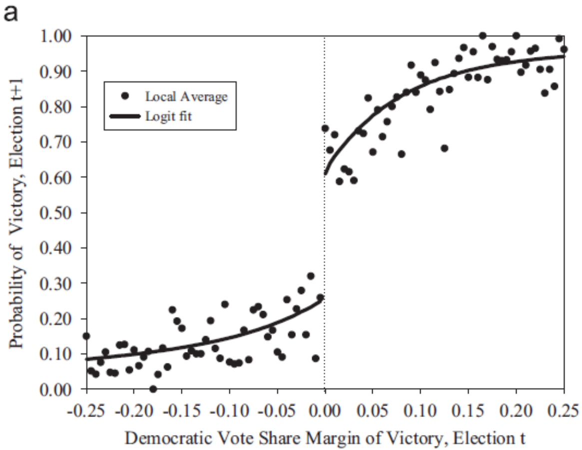
where $\beta_1^* = \beta_{11} - \beta_{01}, \beta_2^* = \beta_{21} - \beta_{21}$ and $\beta_p^* = \beta_{1p} - \beta_{0p}$

- Equation (1) above is a special case of (2) with $\beta_1^* = \beta_2^* = \beta_p^* = 0$.
- The treatment effect at X_o is ρ.
- The treatment effect at $X_i - X_0 = c > 0$ is: $\rho + \beta_1^* c + \beta_2^* c^2 + ... + \beta_p^* c^p$

Example Sharp RD: Lee (2008) Incumbency Effects

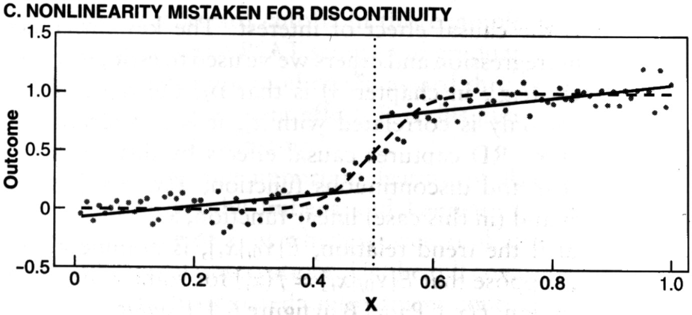
- Lee (2008) uses a sharp RD design to estimate the probability that the incumbent wins an election.
- A large political science literature suggests that incumbents may use privileges and resources of office to gain an advantage over potential challengers.
- An OLS regression of incumbency status on election success is likely to be biased because of unobserved differences. Incumbents have already won an election so they may just be better.
- Lee analyzes the incumbency effect using Democratic incumbents for US congressional elections.
- He analyzes the probability of winning the election in year t+1 by comparing candidates who just won compared to candidates who just lost the election in year t.

The Effect or Winning the Previous Election on The Probability of Winning Current Election



Internal Validity of RD Estimates

- The validity of RD estimates depends crucially on the assumption that the polynomials provide an adequate representation of $E[Y_{0i}|X_i]$.
- If not what looks like a jump may simply be a non-linearity in $f(X_i)$ that the polynomials have not accounted for.



Fuzzy RD

- Fuzzy RD exploits discontinuities in the probability of treatment conditional on a covariate.
- The discontinuity becomes an instrumental variable for treatment status.
- D_i is no longer deterministically related to crossing a threshold but there is a jump in the *probability* of treatment at X_o.

$$P[\mathrm{D}_i = 1|X_i] = \{ \begin{matrix} g_1(X_i) \text{ if } x_i \geq x_o \\ g_0(X_i) \text{ if } x_i < x_o \end{matrix}, \text{ where } g_1(X_i) \neq g_0(X_i)$$

- $g_1(X_i)$ and $g_0(X_i)$ can be anything as long as they differ at x_0.
- The relationship between the probability of treatment and X_i can be written as:

$$P[\mathrm{D}_i = 1|X_i] = g_0(X_i) + [g_1(X_i) - g_0(X_i)]\mathrm{T}_i$$

where $\mathrm{T}_i = 1(X_i \geq X_0)$

Use the Discontinuity as Instrument

- We can write down a first stage relationship:

$$E[\mathrm{D}_i|X_i] = \gamma_{oo} + \gamma_{o1}X_i + \gamma_{o2}X_i^2 + \ ... \ + \gamma_{op}X_i^p$$
$$+ \ \pi \mathrm{T}_i + \gamma_1^* X_i \mathrm{T}_i + \gamma_2^* X_i^2 \mathrm{T}_i + \ ... \ + + \gamma_p^* X_i^p \mathrm{T}_i$$

- One can therefore use both T_i as well as the interaction terms as instruments for D_i.
- If one uses only T_i as IV one has a just identified model which usually has good finite sample properties. In that case the estimated first stage would be:

$$\mathrm{D}_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \ ... \ + \gamma_p X_i^p + \ \pi \mathrm{T}_i + \xi_{1i}$$

- The fuzzy RD reduced form is:

$$\mathrm{Y}_i = \mu + \kappa_1 X_i + \kappa_2 X_i^2 + \ ... \ + \kappa_p X_i^p + \ \rho\pi \mathrm{T}_i + \xi_{2i}$$

Fuzzy RD with Varying Treatment Effects - Second Stage

- As in the sharp RD case one can allow the smooth function to be different on both sides of the discontinuity.
- The second stage model with interaction terms would be the same as before:

$$\begin{aligned} Y_i = {} & \alpha + \beta_{01}\widetilde{x}_i + \beta_{02}\widetilde{x}_i^2 + ... + \beta_{0p}\widetilde{x}_i^p \\ & +\rho D_i + \beta_1^* D_i\widetilde{x}_i + \beta_2^* D_i\widetilde{x}_i^2 + ... + \beta_p^* D_i\widetilde{x}_i^p + \eta_i \end{aligned} \tag{2}$$

- Where $\widetilde{x}$ are now not only normalized with respect to x_o but are also fitted values obtained from the first stage regression.

- Again one can use both T_i as well as the interaction terms as instruments for D_i.
- Only using T the estimated first stages would be:

$$\begin{aligned}\mathrm{D}_i = \gamma_{oo} + \gamma_{o1}\widetilde{X}_i + \gamma_{o2}\widetilde{X}_i^2 + \ldots + \gamma_{op}\widetilde{X}_i^p \\ + \pi \mathrm{T}_i + \gamma_1^* \widetilde{X}_i \mathrm{T}_i + \gamma_2^* \widetilde{X}_i^2 \mathrm{T}_i + \ldots + + \gamma_p^* \widetilde{X}_i^p \mathrm{T}_i + \xi_{1i}\end{aligned}$$

- We would also construct analogous first stages for $\widetilde{X}_i \mathrm{D}_i, \widetilde{X}_i^2 \mathrm{D}_i$, ... ,$\widetilde{X}_i^p \mathrm{D}_i$.

What Does Fuzzy RD Estimate?

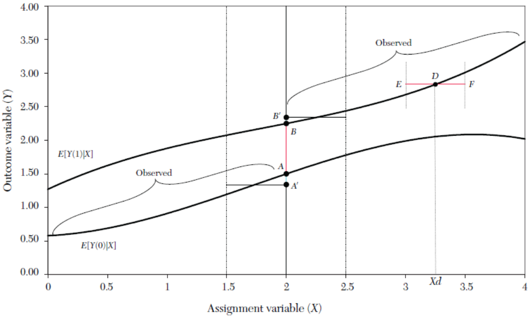
- As Hahn, Todd, and van der Klaauw (2001) point out, one needs the same assumptions as in the standard IV framework.
- As with other binary IVs one then estimates LATE: the average treatment effect of the compliers.
- In RD the compliers are those whose treatment status changes as we move the value of x_i from just the left of x_0 to just to the right of x_0.

Practical Tips for RD Designs: Estimating the f-Function

- As pointed out before there are essentially two ways of approximating the $f(X)$.
    1. Kernel regression.
    2. Using a polynomial function (as outlined above).
- There is no right or wrong method. Both have advantages and disadvantages.

The Kernel Method

- The nonparametric kernel method has its problems in this case because you are trying to estimate regressions at the cutoff point.
- This results in a "boundary problem".



- While the "true" effect is AB, with a certain bandwidth a rectangular kernel would estimate the effect as A'B'.
- There is therefore systematic bias with the kernel method if the $f(X)$ is upwards or downwards sloping.

The Kernel Method - Local Linear Regression

- The standard solution to this problem is to run local linear regression.
- In the case drawn above this would substantially reduce the bias.
- You simply estimate the following model using local linear regression:

$$Y_i = \alpha + \rho D_i + \beta_{01}\widetilde{x}_i + \beta_1^* D_i \widetilde{x}_i + \eta_i$$

where $\widetilde{x}_i = X_i - X_o$

- While estimating this in a given window of width h around the cutoff is straightforward it is more difficult to choose this bandwidth.
- There is essentially a trade-off between bias and efficiency.
- See Lee and Lemieux (2010) for 2 methods to choose the bandwidth.

The Polynomial Method

- Alternatively you can estimate the $f(X)$ function including polynomials in X (see above).
- The polynomial method suffers from the problem that you are using data that is far away from the cutoff to estimate the $f(X)$ function.
- The equivalent of choosing the right bandwidth for the polynomial method is to use the right order of polynomial.
- An easy way suggested by Lee and Lemieux (2010) to test whether you have the right polynomial is to estimate the polynomial function and include a full set of bin dummies in the regression.
  - Then test the null hypothesis whether all bin dummies are 0.
  - Add polynomial terms until you can no longer reject that null hypothesis.

Practical Tips for Estimation

- It is probably advisable to report results for both estimation types:
    1. Polynomials in X.
    2. Local linear regression.
- In robustness checks you also want to show that including higher order polynomials does not substantially affect your findings.
- You also want to show that your results are not affected if you vary the window around the cutoff (standard errors may go up but hopefully the point estimate does not change).

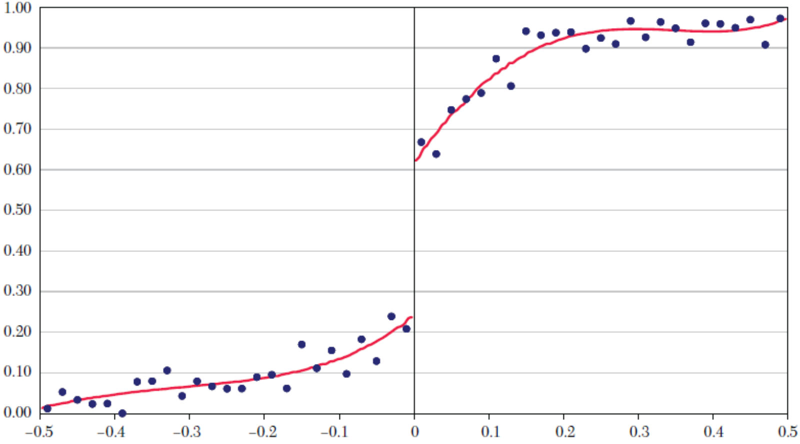
Graphical Analysis in RD Designs

A graphical analysis should be an integral part of any RD study. You should show the following graphs:

1. **Outcome by forcing variable** (X_i) :
   - The standard graph showing the discontinuity in the outcome variable.
   - Construct bins and average the outcome within bins on both sides of the cutoff.
   - You should look at different bin sizes when constructing these graphs (see Lee and Lemieux (2010) for details).
   - Plot the forcing variable X_i on the horizontal axis and the average of Y_i for each bin on the vertical axis.
   - You may also want to plot a relatively flexible regression line on top of the bin means.
   - Inspect whether there is a discontinuity at x_0.
   - Inspect whether there are other unexpected discontinuities.

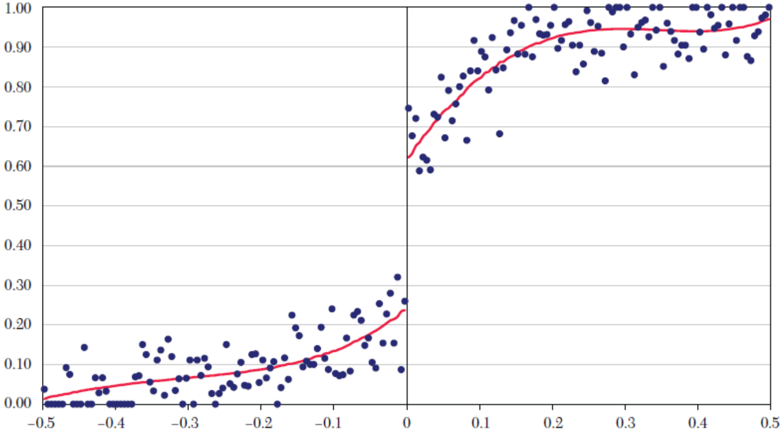
Example: Outcomes by Forcing Variable

From Lee and Lemieux (2010) based on Lee (2008)



Example: Outcomes by Forcing Variable - Smaller Bins

From Lee and Lemieux (2010) based on Lee (2008)



2. **Probability of treatment by forcing variable if fuzzy RD.**
   - In a fuzzy RD design you also want to see that the treatment variable jumps at $x_{0.}$
   - This tells you whether you have a first stage.
3. **Covariates by forcing variable.**
   - Construct a similar graph to the one before but using a covariate as the "outcome".
   - There should be no jump in other covariates.
   - If the covariates would jump at the discontinuity one would doubt the identifying assumption.

Example Covariates by Forcing Variable

From Lee and Lemieux (2010) based on Lee (2008)

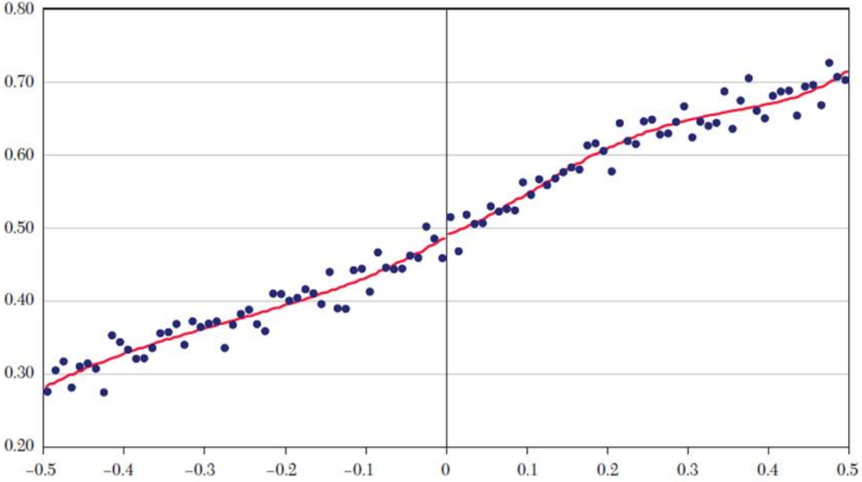


Figure 17. Discontinuity in Baseline Covariate (Share of Vote in Prior Election)

4. **The density of the forcing variable.**
   - One should plot the number of observations in each bin.
   - This plot allows to investigate whether there is a discontinuity in the distribution of the forcing variable at the threshold.
   - This would suggest that people can manipulate the forcing variable around the threshold.
   - This is an indirect test of the identifying assumption that each individual has imprecise control over the assignment variable.

Density of the forcing variable

From Lee & Lemieux (2010) based on Lee (2008)

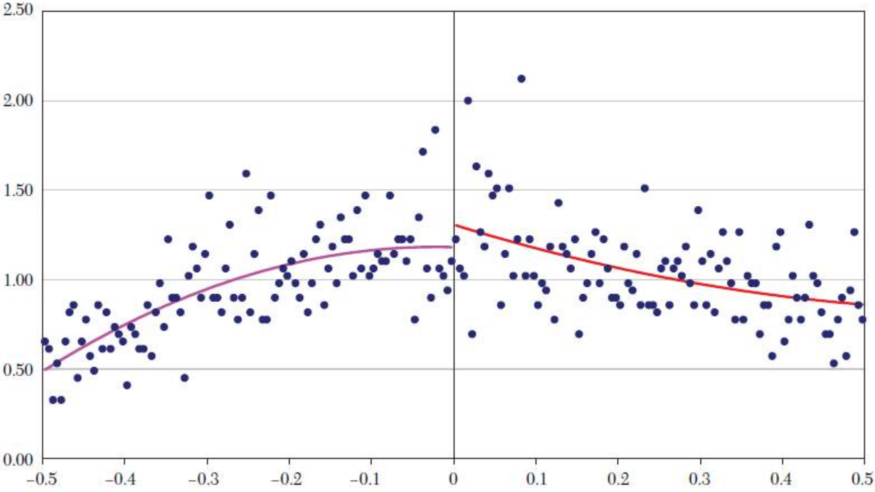


Figure 16. Density of the Forcing Variable (Vote Share in Previous Election)

Testing the Validity of the RD Design

- As outlined above the key identifying assumption is that $E[Y_{0i}|X_i]$ and $E[Y_{1i}|X_i]$ are continuos in X_i at X_0.
- This implies that each individual has imprecise control over the assignment variable.
- It is impossible to test this directly but we can nonetheless get some evidence with the following specification tests.

1. **Testing the continuity of the density of X:**
   - McCrary (2008) suggests testing the null hypothesis of continuity of the density of the forcing variable at the discontinuity point.
   - In principle one does not need continuity. A discontinuity in the density, however, suggests that there is some manipulation of X around the threshold going on.
   - In the first step you partition the assignment variable into bins and calculate frequencies (number of observations) in the bins.
   - In the second step you treat those frequency counts as dependent variable in a local linear regression as before.
   - McCrary adopts the nonparametric framework for asymptotics.
   - See his website (http://www.econ.berkeley.edu/~jmccrary/DCdensity/) for details on the test.

Testing the Validity of the RD Design

2. **Test involving covariates:**
   - Test whether other covariates exhibit a jump at the discontinuity. (Just re-estimate the RD model with the covariate as the dependent variable).
   - This is a type of placebo test.
3. **Testing for jumps at non-discontinuity points:**
   - Imbens and Lemieux (2008) suggest to only look at one side of the discontinuity and take the median of the forcing variable in that section and test whether you can find a discontinuity in that part.
   - Another type of placebo test.

An Application of Fuzzy RD on Class Sizes

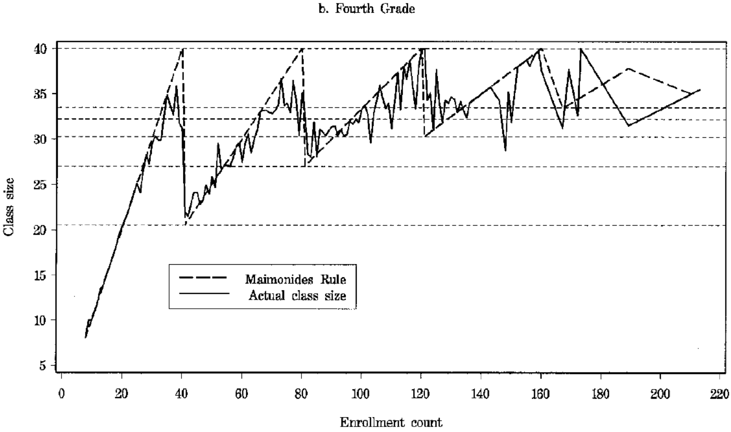
- Angrist and Lavy (1999) use a fuzzy RD design to analyze the effect of class size on test scores.
- They extend RD in two ways compared to the discussion above:
  1. The causal variable of interest (class size) takes on many values. $\rightarrow$ the first stage exploits discontinuities in average class size instead of probabilities of a single treatment.
  2. They use multiple discontinuities.
- Angrist and Lavy exploit an old Talmudic rule that classes should be split if they have more than 40 students in Israel.
  - A school with 40 students has only one class. $\rightarrow$ class size 40.
  - A school with 41 students has two classes. $\rightarrow$ class sizes 21 and 20.
- Predicted class size from a strict application of Maimonides rule is:

$$m_{sc} = \frac{e_s}{int[\frac{(e_s-1)}{40}]+1}$$

where $int[a]$ is the integer part of real number a.

e_s is enrollment

Maimonides Rule and Actual Class Size



- The rule is not followed completely strictly they therefore have a fuzzy discontinuity design.

Econometric Specification

- They want to estimate the relationship between average achievement and class size.

$$Y_{isc} = \alpha_0 + \rho n_{sc} + \eta_{isc}$$

- Estimating this relationship with OLS may lead to biased results because class size is likely to be correlated with the error term. The 2 main reasons for this are:
  1. Parents from higher socioeconomic backgrounds may put their children in schools with smaller classes.
  2. Because principals may put weaker students in smaller classes.

Fuzzy RD Design

- Angrist & Lavy therefore use the Maimonides rule in a fuzzy RD design.

$$\mathrm{Y}_{isc} = \alpha_0 + \alpha_1 d_s + \rho n_{sc} + \beta_1 e_s + \beta_2 e_s^2 + \eta_{isc} \qquad (1)$$

  where Y_{isc} is the test score of student i in school s and class c.
  e_s is enrollment in school s.
  d_s is the percentage of disadvantage students in class
  n_{sc} is class size.

- The variables relate to the previous description as follows:
  - n_{sc} plays the role of D_i.
  - e_s plays the role of X_i.
  - m_{sc} plays the role of T_i.
- The first stage regression is:

$$n_{sc} = \gamma_0 + \gamma_1 d_s + \pi m_{sc} + \delta_1 e_s + \delta_2 e_s^2 + \xi_{isc} \qquad (2)$$

  where m_{sc} is the function describing Maimonides rule.

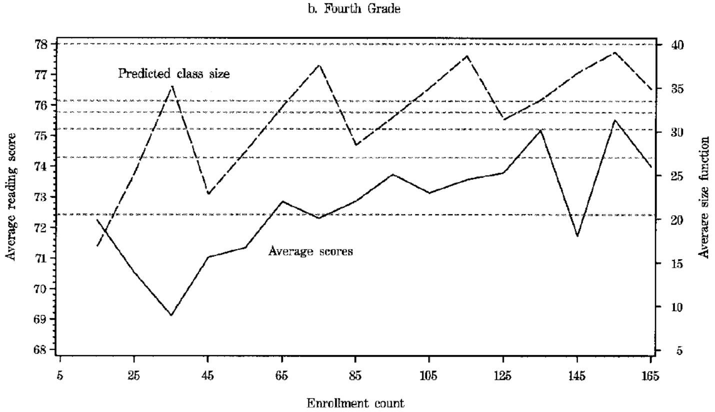
Regression Results - OLS (5th Grade)

	5th Grade					
	Reading comprehension			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
Mean score		74.3			67.3	
(s.d.)		(8.1)			(9.9)	
Regressors						
Class size	.221	−.031	−.025	.322	.076	.019
	(.031)	(.026)	(.031)	(.039)	(.036)	(.044)
Percent disadvantaged		−.350	−.351		−.340	−.332
		(.012)	(.013)		(.018)	(.018)
Enrollment			−.002			.017
			(.006)			(.009)
Root MSE	7.54	6.10	6.10	9.36	8.32	8.30
R^2	.036	.369	.369	.048	.249	.252
N		2,019			2,018	

- There is a positive OLS relationship between class size and test scores. If you control for percentage disadvantaged and total enrollment, however, the relationship turns slightly negative but not significantly so.

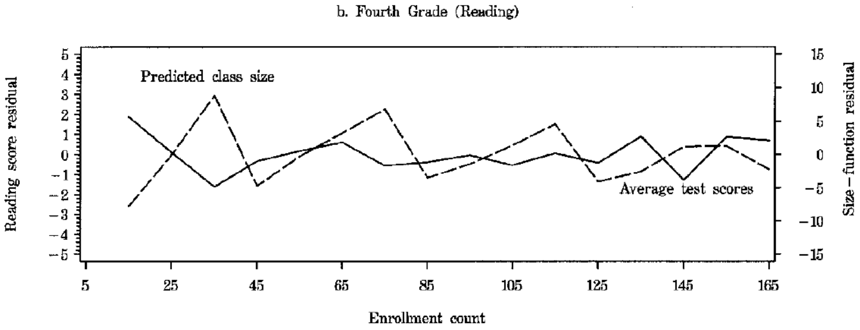
Reduced Form Relationship of Maimonides Rule on Test Scores

- Students in schools with more overall enrollment (often in bigger cities) do better on average.
- Average test scores are partly a mirror image of predicted class sizes.



Reduced Form Relationship of Maimonides Rule on Test Scores

- Because larger schools are often in better-off areas they control for enrolment when they redraw the relationship between class-size and achievement.
- Now test-scores are more of a mirror image to predicted class sizes.



RD Second Stage Reading - 5th Graders

	Reading comprehens			
	Full sample			
	(1)	(2)	(3)	(4)
Mean score		74.4		
(s.d.)		(7.7)		
Regressors				
Class size	−.158	−.275	−.260	−.186
	(.040)	(.066)	(.081)	(.104)
Percent disadvantaged	−.372	−.369	−.369	
	(.014)	(.014)	(.013)	
Enrollment		.022	.012	
		(.009)	(.026)	
Enrollment squared/100			.005	
			(.011)	
Piecewise linear trend				.136
				(.032)
Root MSE	6.15	6.23	6.22	7.71
N		2019		1961

- The effect of class size now is significantly negative.

RD Second Stage Reading - Discontinuity Sample

	Reading comprehension					
	Full sample				+/− 5 Discontinuity sample	
	(1)	(2)	(3)	(4)	(5)	(6)
Mean score		74.4			74.5	
(s.d.)		(7.7)			(8.2)	
Regressors						
Class size	−.158	−.275	−.260	−.186	−.410	−.582
	(.040)	(.066)	(.081)	(.104)	(.113)	(.181)
Percent disadvantaged	−.372	−.369	−.369		−.477	−.461
	(.014)	(.014)	(.013)		(.037)	(.037)
Enrollment		.022	.012			.053
		(.009)	(.026)			(.028)
Enrollment squared/100			.005			
			(.011)			
Piecewise linear trend				.136		
				(.032)		
Root MSE	6.15	6.23	6.22	7.71	6.79	7.15
N		2019		1961	471	

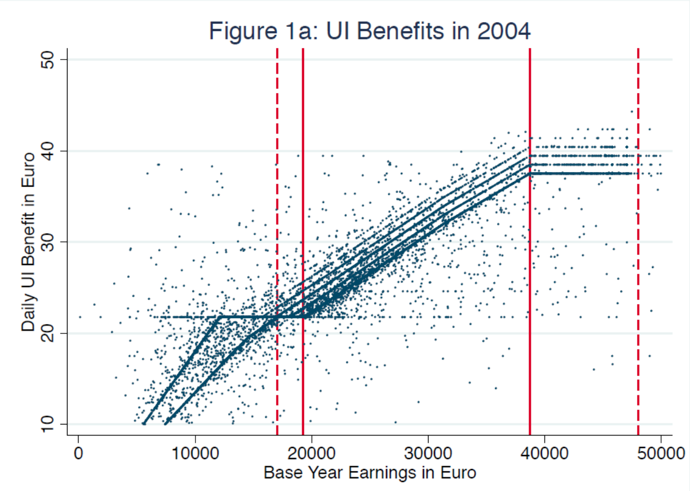
- Coefficients in discontinuity sample are fairly similar (for math they are even more similar).

- Card, Lee, Pei and Weber (2012) introduce a variant of the RD design which they call regression kink design (RKD).
- They essentially use a kink in some policy rule to identify the causal effect of the policy.
- Instead of a jump in the outcome you now expect a jump in the first derivative.

Unemployment Benefits in Austria

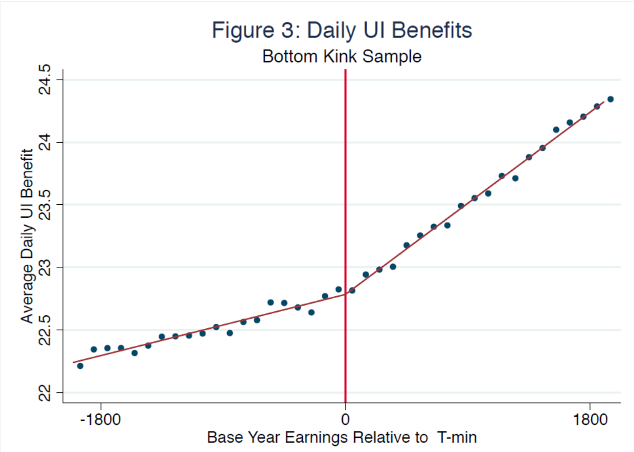
- They apply their design to answer the question whether the level of unemployment benefits affects the length of unemployment in Austria.
- Unemployment benefits are based on income in a base period.
- The benefit formula for unemployment exhibits 2 kinks.
  - There is a minimum benefit level (that is not binding for people with very low earnings)
  - Then benefits are 55% of the earnings in the base period
  - There is a maximum benefit level that is adjusted every year
- People with dependents get small supplements (that is why one can distinguish five "solid" lines in the following graph).
- Not everyone receives benefits that correspond one to one to the formula because of mistakes in the administrative data.

Base Year Earnings and Unemployment Benefits



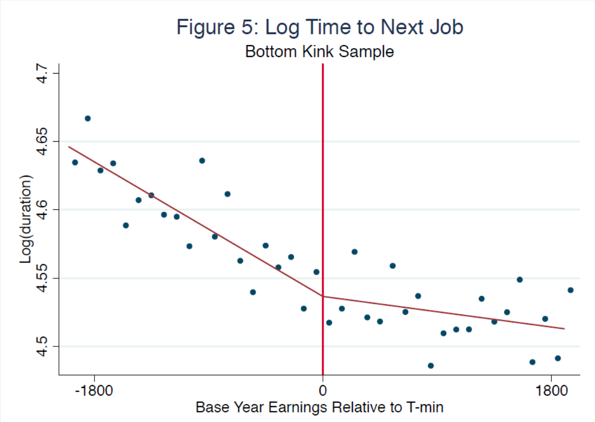
The graph shows unemployment benefits (vertical axis) as a function of pre-unemployment earnings (horizontal axis).

Base Year Earnings and Benefits for Single Individuals



- Bin-Size: 100 Euros
- For single individuals UI benefits are flat below the cutoff. The relationship is still upward sloping because of family benefits.

Time to Next Job For Single Individuals



- People with higher base earnings have less trouble finding a job (negative slope).
- There is a kink: the relationship becomes shallower once benefits increase more.