4. Regression Discontinuity Designs

Econometrics II
Winter 2019
Osaka U

Shuhei Kitamura

Summary of the last topic

Imperfect compliance

Local average treatment effect (LATE)

Intention to treat (ITT) effect

Two-stage least squares (2SLS)

IV assumptions

IV solves measurement error

Outline

Shape RD

Fuzzy RD

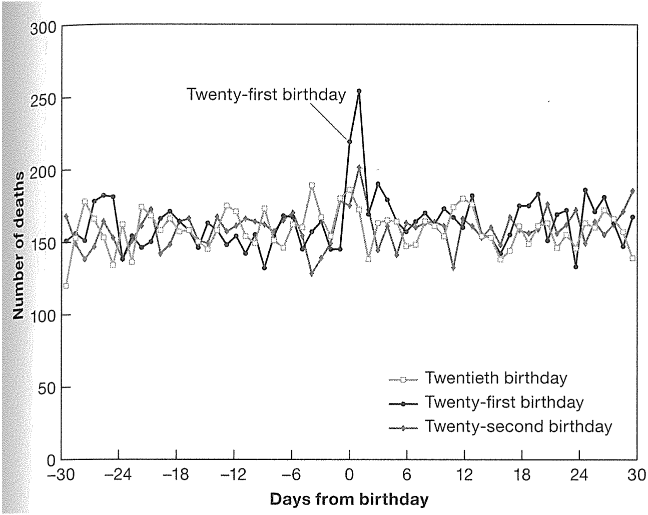
Sharpe RD: An example

In the U.S., the minimum legal drinking age (MLDA) is 21.

- Your life might be quite different just a day before/after your 21st birthday.

There is no complier issue because it is determined by law.

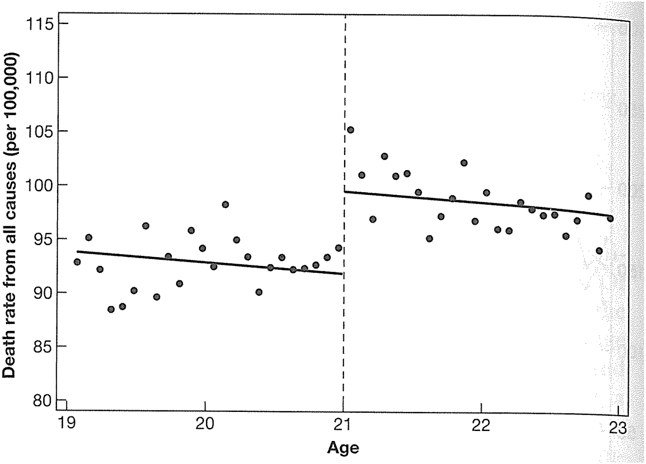
Birthdays and deaths



Source: Metrics.

Notes: The number of deaths among Americans aged 20-22 between 1997-2003.

Birthdays and deaths (cont.)



Source: Metrics.

MLDA

You see in the figure that death rates jump at the age-21 cutoff (but not at other age cutoffs).

- If there was no rule like the MLDA, we might have not seen any jump at the cutoff.

RD designs use this kind of jump in the trend line.

MLDA (cont.)

With mathematical notations,

$$D_a = \begin{cases} 1 & \text{if } a \geq a_0 \\ 0 & \text{if } a < a_0, \end{cases} \tag{1}$$

Where a_0 is the cutoff.

In the MLDA example, the running variable is age, which satisfies

- Treatment status is a deterministic function of age. (Once we know age, we know D_a.)
- Treatment status is a discontinuous function of age. (No matter how age gets close to the cutoff, D_a does not change until the cutoff is reached.)

In sharp RD, treatment status jumps when the running variable passes the cutoff.

Causal effect in Sharp RD

ATE is

$$\beta = E[Y_i(1) - Y_i(0) | X_i = c], \tag{2}$$

Where c is the cutoff. ATE is identified as

$$\lim_{x \downarrow c} E[Y_i | X_i = x] - \lim_{x \uparrow c} E[Y_i | X_i = x]. \tag{3}$$

What kind of comparison are we making here?

- Treatment group? Control group?
- Common support?
- RDD relies on our willingness to extrapolate across values of the running variable, at least in a neighborhood of the cutoff.

RDD relies on the assumption that individuals very close to the cutoff are likely to be very similar.

- Looking at the individuals near the cutoff, we can imitate the random assignment. Individuals to the right of the cutoff are treated, while individuals to the left of the cutoff are control.

MLDA (cont.)

RDD uses a regression like

$$M_a = \alpha + \beta D_a + \gamma a + \varepsilon_a, \tag{4}$$

Where M_a is the death rate in month a and a is age in month (month is defined as a 30-day interval counting from the 21st birthday).

- One increment is about 0.08 years old (1/365 * 30)

This γa captures a linear trend in case there is no jump.

Thus the equation (4) works if the trend is linear.

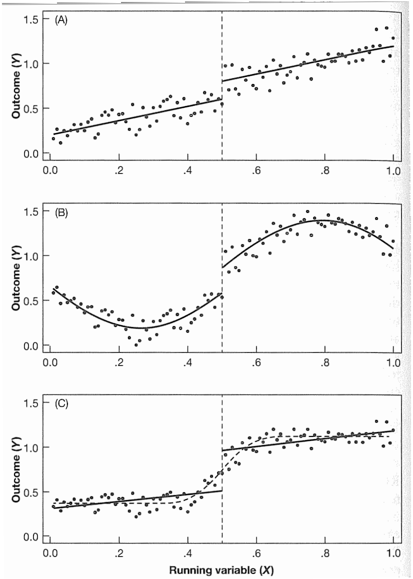
By contrast, if there is nonlinearity in the trend curve around the cutoff, the above model may mistakenly interpret it as discontinuity.

MLDA (cont.)

There are two ways to alleviate the issue

- Model nonlinearity (parametric)
- Use observations very near the cutoff (nonparametric)

Model nonlinearity



Absorb nonlinearity

One way to model nonlinear trends is to add a quadratic control

$$M_a = \alpha + \beta D_a + \gamma_0 a + \gamma_1 a^2 + \varepsilon_a. \tag{5}$$

What shape of trend does this control absorb?

Allow slope change

One way to model a slope change around the cutoff is

$$M_a = \alpha + \beta D_a + \gamma(a - a_0) + \lambda (a - a_0) D_a + \varepsilon_a, \tag{6}$$

Where a_0 is the cutoff ($= 21$ in our example).

By the way, why not

$$M_a = \alpha + \beta D_a + \gamma a + \lambda a D_a + \varepsilon_a \tag{7}$$

?

Let's see why the first model is the correct formulation.

Model nonlinearity (cont.)

The model allows different coefficients to the left and the right of the cutoff. Ignoring the constant, one can show

$$\begin{cases} \beta + (\gamma + \lambda)(a - a_0) & \text{if } a \text{ is above the cutoff} \\ \beta & \text{if } a \text{ is at the cutoff} \\ \gamma(a - a_0) & \text{if } a \text{ is below the cutoff} \end{cases} \tag{8}$$

The treatment effect away from the cutoff is $\beta + \lambda(a - a_0)$.

- It changes as a changes.
- It may be informative but as well be speculative.

By contrast, the effect at the cutoff is not affected by it.

Model nonlinearity (cont.)

Or you can combine models dealing with nonlinear trends and a change in slope

$$M_a = \alpha + \beta D_a + \gamma_0 \bar{a} + \gamma_1 \bar{a}^2 + \lambda_0 \bar{a} D_a + \lambda_1 \bar{a}^2 D_a + \varepsilon_a, \qquad (9)$$

Where $\bar{a} = a - a_0$.

Model nonlinearity (cont.)

How much nonlinearity do we need?

- Is quadratic control enough?

It requires a judgment call.

- Complex models are not always better than a simple model.
- Report results with several different specifications. For example, report the simplest (4) and a more complicated (9).

You should also plot data and fitted values from regressions for several variables including outcomes and controls.

R exercise

Launch RStudio.

Type

```
rd_mlda <- mmdata::rd_mlda
rd_mlda <- subset(rd_mlda,(is.na(all)==FALSE))

rd_mlda$agecell2 <- rd_mlda$agecell^2

rd_mlda$age <- rd_mlda$agecell - 21
rd_mlda$age2 <- rd_mlda$age^2

rd_mlda$over21 <- ifelse(rd_mlda$agecell >= 21,1,0)

rd_mlda$over_agecell <- rd_mlda$over21 * rd_mlda$agecell

rd_mlda$over_age <- rd_mlda$over21 * rd_mlda$age
rd_mlda$over_age2 <- rd_mlda$over_age^2
```

R exercise (cont.)

Type

```
reg1 <- lm(all ~ over21 + agecell, data=rd_mlda)
stargazer(reg1, type="text")
```

This is the baseline regression

$$M_a = \alpha + \beta D_a + \gamma a + \varepsilon_a. \tag{10}$$

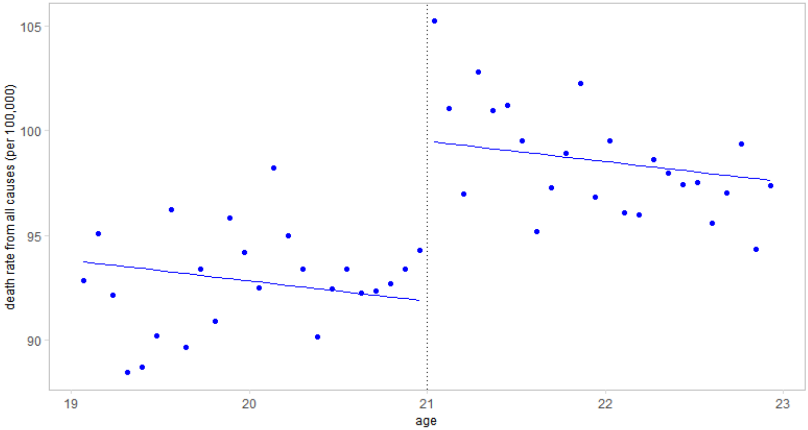
R exercise (cont.)

Let's plot it.

Type

```
rd_mlda$pred_all <- fitted(reg1)

ggplot(data=rd_mlda, aes(x=agecell, y=all)) +
    geom_point() +
    geom_line(aes(x=agecell, y=pred_all)) +
    geom_vline(xintercept=21, linetype='dotted')
```

Linear trend



R exercise (cont.)

Let's add a quadratic term

```
reg2 <- lm(all ~ over21 + agecell + agecell2,
data=rd_mlda)
stargazer(reg1, reg2, type="text")
```

R exercise (cont.)

Let's add change in slope

```
reg3 <- lm(all ~ over21 + agecell + over_agecell,
data=rd_mlda)
reg4 <- lm(all ~ over21 + age + over_age, data=rd_mlda)
stargazer(reg1, reg2, reg3, reg4, type="text")
```

Comparing the estimates of `over21` in `reg3` and `reg4`, what do you find?

R exercise (cont.)

Let's add both of them

```
reg5 <- lm(all ~ over21 + age + age2 + over_age +
over_age2, data=rd_mlda)
stargazer(reg1, reg2, reg3, reg4, reg5, type="text")
```

Comparing the estimate of `over21` in `reg5` with the other estimates of `over21`, what do you find?

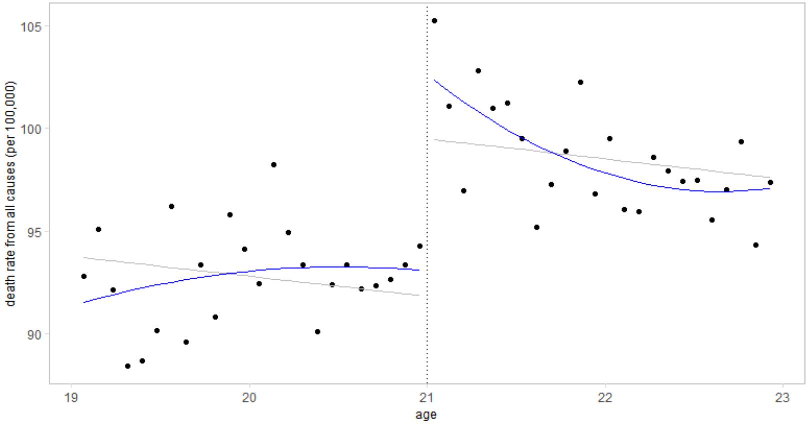
R exercise (cont.)

Let's plot it.

Type

```
rd_mlda$pred_all_q <- fitted(reg5)

ggplot(data = rd_mlda, aes(x=agecell, y=all)) +
    geom_point() +
    geom_line(aes(x=agecell, y=pred_all), color='gray') +
    geom_line(aes(x=agecell, y=pred_all_q)) +
    geom_vline(xintercept = 21, linetype='dotted')
```

Linear trend vs. quadratic trend with different slopes



Limit sample

The second strategy is to focus only on data points close to a cutoff. That is, you estimate

$$M_a = \alpha + \beta D_a + \gamma a + \varepsilon_a, \tag{11}$$

For a narrow window $a_0 - b \leq a \leq a_0 + b$, where a_0 is the cutoff and b is a bandwidth.

A drawback is that the sample size gets smaller as the window size gets narrower.

R exercise

Let's limit sample to 20-22 year olds in the MLDA example.

- The original data include individuals between 19.07-22.93 years old.

Type

```
rd_mlda_limit <- subset(rd_mlda, (agecell >= 20 & agecell
<= 22))

reg1 <- lm(all ~ over21 + agecell, data=rd_mlda)
reg2 <- lm(all ~ over21 + age + age2 + over_age +
over_age2, data=rd_mlda)
reg3 <- lm(all ~ over21 + agecell, data=rd_mlda_limit)
reg4 <- lm(all ~ over21 + age + age2 + over_age +
over_age2, data=rd_mlda_limit)

stargazer(reg1, reg2, reg3, reg4, type="text")
```

What do you find?

Limit sample (cont.)

How can we determine the bandwidth?

Again, this requires a judgment call.

- Report how results are sensitive to different bandwidths.

There is also a bandwidth selection algorithm.

- Imbens and Kalyanaraman (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79 (3), 933-959.
- `IKbandwidth` function in the `rdd` package computes this for you.

Placebo test

If we believe that the effect we observe is due to the MLDA

- We should observe the effect for external causes such as motor vehicle accidents (mva)
- BUT we should not observe the effect for any internal causes such as cancer or other diseases

Let's check them using the data.

This kind of exercise is often called a placebo test.

R exercise

Type

```
reg1 <- lm(internal ~ over21 + age + age2 + over_age +
over_age2, data=rd_mlda)
reg2 <- lm(mva ~ over21 + age + age2 + over_age +
over_age2, data=rd_mlda)

stargazer(reg1, reg2, type="text")
```

R exercise (cont.)

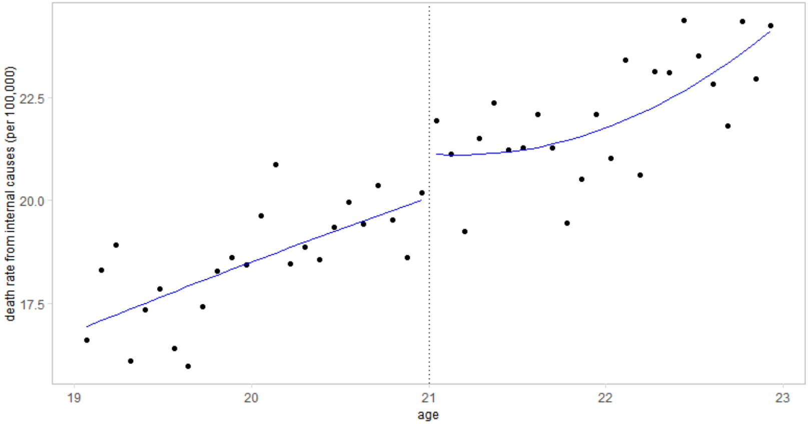
Type

```
ggplot(data=rd_mlda, aes(x=agecell, y=internal)) +
    geom_point() +
    geom_line(aes(x=agecell, y=pred_internal)) +
    geom_vline(xintercept = 21, linetype="dotted")
```

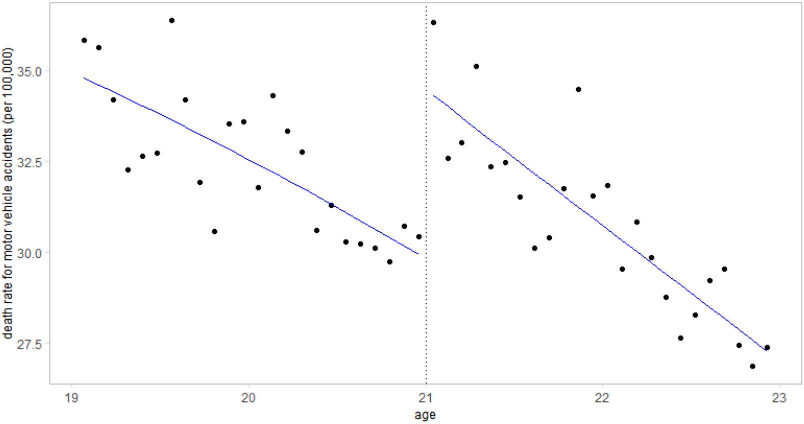
Then

```
ggplot(data=rd_mlda, aes(x=agecell, y=mva)) +
    geom_point() +
    geom_line(aes(x=agecell, y=pred_mva)) +
    geom_vline(xintercept = 21, linetype="dotted")
```

Internal causes



Auto vehicle accidents



Islamic rule and women's empowerment

Does Islamic political control affect women's empowerment?

- A often quoted concern is that Islamic political control will endanger gender equality.
- But is there any endogeneity issue?

Islamic rule and women's empowerment (cont.)

Turkey experienced a political change in the local elections in 1994.

- The pro-Islamic *Refah* Party (Welfare Party) became the third largest party in terms of voting.
- The party was later banned by the Constitutional Court because it violated the separation of religion and state.

What is the effect of the winning of the Islamic party on later educational outcomes of women in the constituencies?

Islamic rule and women's empowerment (cont.)

Meyerson (2014) tackles the question by using an RDD.

Consider the following model

$$Y_i = \alpha + \beta D_i + f(X_i) + \varepsilon_i \tag{12}$$

Where D_i is the treatment, and X_i is the forcing variable.

- X_i is the win margin for the Islamic party relative to the largest non-Islamic party.
- D_i takes value one if X_i is larger or equal to the cutoff $(= 0)$.

The bandwidth is determined by using the Imbens-Kalyanaraman algorithm.

Main outcome: The share of 15-20 year-old females who have completed high school by 2000, according to the 2000 census.

R exercise

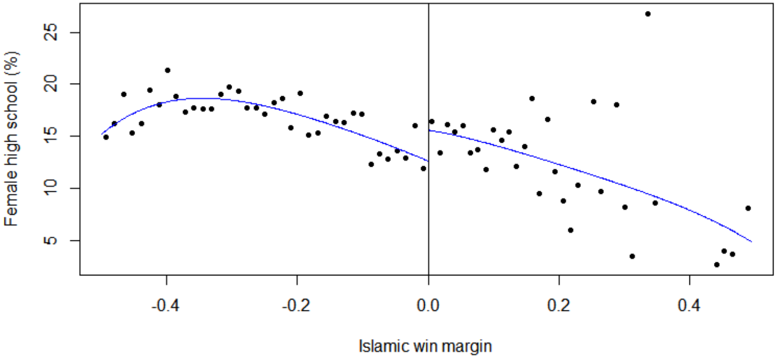
Launch RStudio.

Type

```
setwd("path_to_your_folder")

turkey <- read.csv("turkey.csv")

res <- compareGroups(T ~ Y, data=turkey)
createTable(res)

reg1 <- lm(Y T̃, data=turkey)
stargazer(reg1, type=''text'')
```

What do you find?

Discontinuity



- Loot, the figure indicates the opposite sign near the cutoff!

R exercise (cont.)

Type

```
turkey$X2 <- turkey$X^2
turkey$T_X <- turkey$T * turkey$X
turkey$T_X2 <- turkey$T * turkey$X2

bw_ik <- IKbandwidth(turkey$X, turkey$Y, cutpoint=0)

reg2 <- lm(Y ~ T + X + T_X, data=turkey, subset=(X >=
-bw_ik & X <= bw_ik))

stargazer(reg1, reg2, type="text")
```

What do you find?

Table II from Meyerson (2014)

TABLE II

ISLAMIC RULE AND HIGH SCHOOL EDUCATION[a]

Outcome	Completed High School in 2000								Enrollment
Age Cohort	15–20								15–30
Control Function	None		Linear				Quadratic	Cubic	Linear
Bandwidth	Global		$\hat{h}$		$\hat{h}/2$	$2\hat{h}$	$\hat{h}$	$\hat{h}$	$\hat{h}$
Covariates	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				Panel A: Women					
Outcome mean	0.163	0.163	0.152	0.152	0.144	0.166	0.152	0.152	0.127
Islamic mayor in 1994	−0.026***	0.012**	0.032***	0.028***	0.032***	0.022***	0.028***	0.043***	0.014***
	(0.006)	(0.006)	(0.010)	(0.007)	(0.011)	(0.006)	(0.011)	(0.016)	(0.005)
Bandwidth	1.000	1.000	0.240	0.240	0.120	0.480	0.240	0.240	0.205
R^2	0.01	0.55	0.03	0.65	0.65	0.58	0.65	0.65	0.48
Observations	2629	2629	1020	1020	589	2049	1020	1020	904

Density test

If the cutoff is randomly assigned, there should not be any "adjustment" around the cutoff.

Instead, if you find a discontinuity in density around the cutoff, your RDD estimate is likely to be biased.

A test, proposed by McCrary (2008), checks a discontinuity in the density of the running variable around the cutoff.

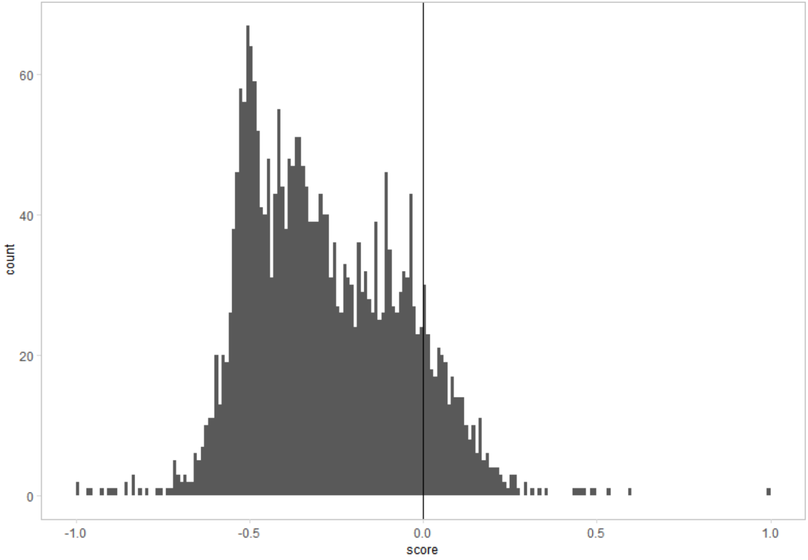
- `DCdensity` in the `rdd` package does this for you.

R exercise (cont.)

Type

```
temp <- as.data.frame(turkey$X)
colnames(temp) <- c("var")

ggplot(data=temp,aes(x=var, y=..count..))  +
    geom_histogram(breaks=seq(-1, 1, 0.01)) +
    geom_vline(xintercept=0) +
    labs(x="score", y="count")
```

Histogram

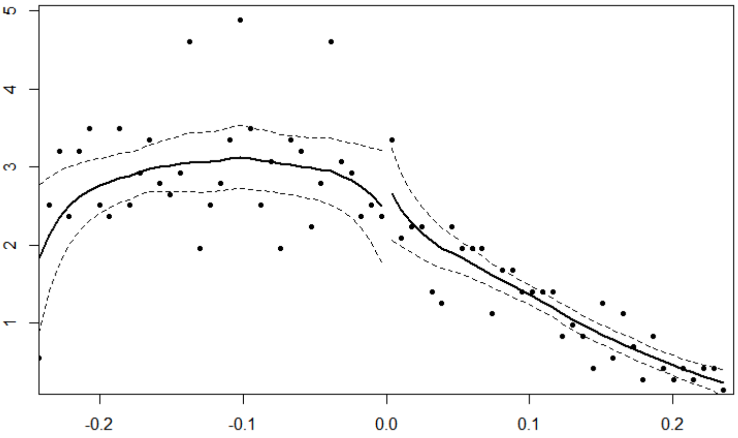


R exercise (cont.)

Type

```
DCdensity(turkey$X, cutpoint=0)
```

Density test

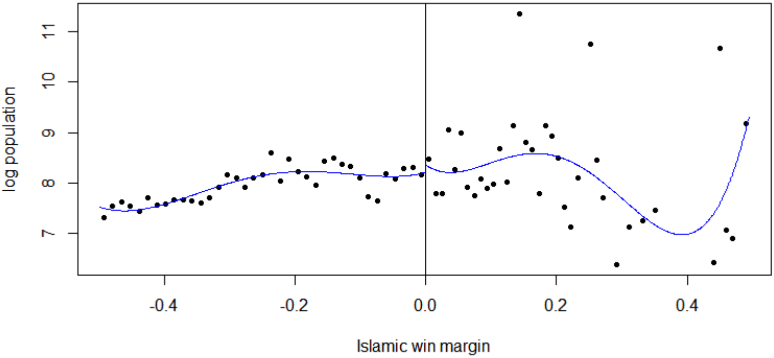


Check other variables

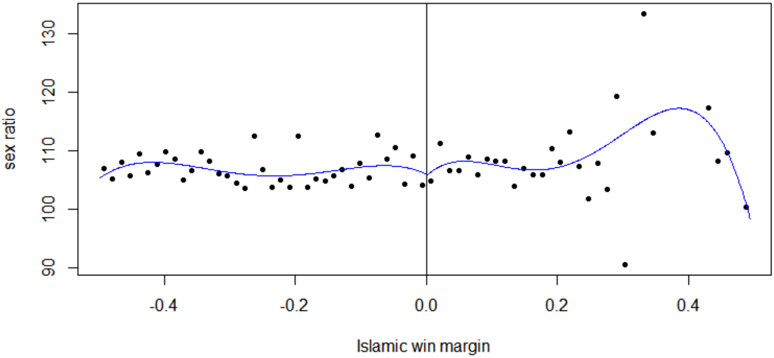
Is there any discontinuity in other variables?

Let's check them.

Population



Sex ratio



Vote share of Islamic party

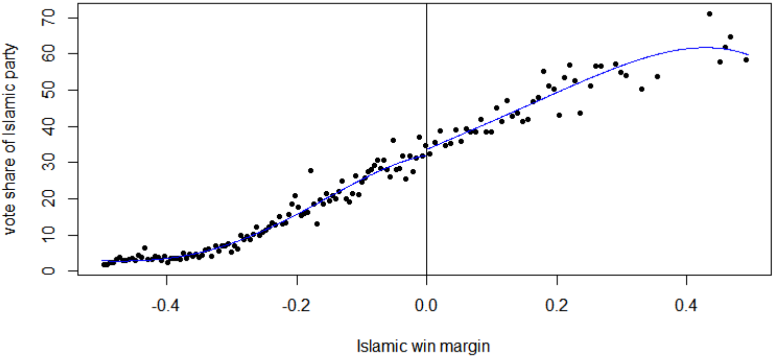
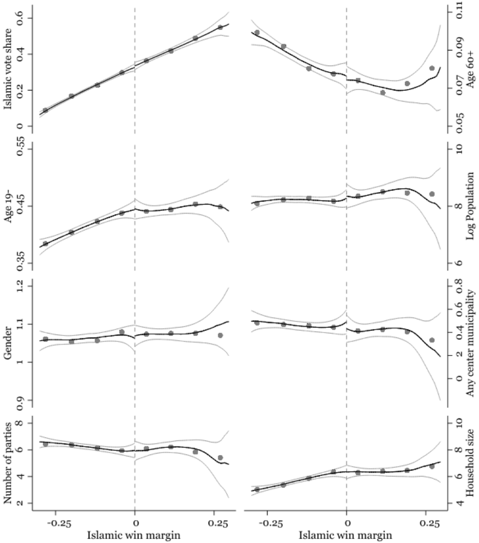


Figure 3 from Meyerson (2014)



Minimum rules for empiricists: Sharp RD

Show visually

- 1. A jump at the cutoff for outcome variables
- 2. No jump for variables which should not jump at the cutoff
- 3. No jump for the density of the running variable at the cutoff

Show the results with different specifications and bandwidths

Fuzzy RD

The difference between Sharp RD and Fuzzy RD is the following

- Sharp RD: There is a clear "on and off" at the cutoff

$$\text{Prob}[D_i = 1 | X_i \geq c] = 1, \tag{13}$$

$$\text{Prob}[D_i = 1 | X_i < c] = 0. \tag{14}$$

- Fuzzy RD: Exploits the discontinuity in the *probability*

$$\lim_{x \downarrow c} \text{Prob}[D_i = 1 | X_i = x] \neq \lim_{x \uparrow c} \text{Prob}[D_i = 1 | X_i = x]. \tag{15}$$

Peer effects in education

Suppose that you are interested in estimating peer effects in school.

Among the applicants to the top school, you could compare individuals who passed its entrance exam with those who did not (and went to another school).

- Comparing those individuals at the cutoff of the entrance exam, you can possibly estimate the causal effect of peer effects in school life on educational achievement (e.g., test scores).

Boston Latin School

There are several exam schools in Boston and New York City.

- The top school among them in Boston is Boston Latin School (BLS).

A research design is to compare school outcomes for individuals who passed the entrance exam and for those who did not, among applicants to BLS.

In particular, we ask

- Are there any peer effects?
- Are math scores affected by peer effects?

Idea: Entering BLS might increase the exposure to high peer quality, yielding better educational outcomes.

We employ Fuzzy RD.

- Why not Sharp RD?

Causal effect in Fuzzy RD

Fuzzy RD is an IV.

- It captures the causal effect on compliers.
- Who are compliers in this case?

LATE is written by

$$E[Y_i(1) - Y_i(0) | C_i = 1, X_i = c], \tag{16}$$

Where $C_i = 1$ indicates complier.

LATE is identified by

$$\frac{\lim_{x \downarrow c} E[Y_i | X_i = x] - \lim_{x \uparrow c} E[Y_i | X_i = x]}{\lim_{x \downarrow c} E[D_i | X_i = x] - \lim_{x \uparrow c} E[D_i | X_i = x]}. \tag{17}$$

The Fuzzy RD estimand boils down to the Sharp RD estimand if the denominator is one.

Boston Latin School (cont.)

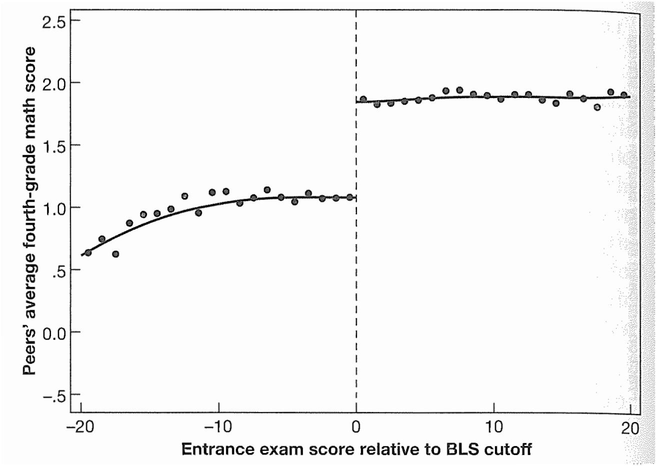
The first-stage regression is

$$P_i = \alpha + \beta D_i + \gamma R_i + \varepsilon_i, \tag{18}$$

Where P_i is peer quality as measured by the average fourth-grade math score of i's seventh-grade classmate, D_i is a dummy taking the value one if i is assigned a seat, and zero otherwise, and R_i is the running variable that determines D_i (i.e., the entrance exam score).

β is estimated as $.8\sigma$ in the example.

Peer's average score



Source: Metrics.

Boston Latin School (cont.)

The second-stage regression is

$$Y_i = a + \rho \hat{P}_i + \delta R_i + u_i \tag{19}$$

Where Y_i is i's seventh-grade math score and $\hat{P}_i$ is the first-stage fitted value.

ρ is estimated as -0.23 with the standard error of $.132$.

- We do not find a causal effect of peer quality.
- Are the IV assumptions likely to be satisfied?

Minimum rules for empiricists: Fuzzy RD

The same as those for IV and Sharp RD (hence omitted).

Summary

Shape RD

Fuzzy RD